IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Attorney Docket No.:	**DEX-0172**
Inventors:	**Salceda et al.**
Serial No.:	**09/763,978**
Filing Date:	**April 25, 2001**
Examiner:	**Aeder, Sean E.**
Customer No.:	**32800**
Group Art Unit:	**1642**
Confirmation No.:	**3638**
Title:	**A Novel Method of Diagnosing, Monitoring, Staging, Imaging and Treating Various Cancers**

"Express Mail" Label No. EM182570636US
Date of Deposit: January 21, 2010

I hereby certify that this paper is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 CFR 1.10 on the date indicated above and is addressed to the Commissioner for Patents, Mail Stop AF, P. O. Box 1450, Alexandria, VA 22313-1450.

By Kathleen A. Tyrrell
Typed Name: Kathleen A. Tyrrell, Reg. No. 38,350

Mail Stop AF
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

APPEAL BRIEF

TABLE OF CONTENTS

I. Real Party in Interest

The real party in interest of this Appeal is the Assignee of the above-referenced patent application, diaDexus, Inc.

II. **Related Appeals and Interferences**

The appellant, the appellant's legal representative, and the assignees are not aware of any other appeals or interferences which will directly affect or be directly affected by or have a bearing on the Board's decision in the instant appeal.

III. **Status of the Claims**

Claims 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 and 13 are canceled.

Claim 14 is rejected and is the subject of appeal.

Claims 15, 16, 17, 18, 19 and 20 are canceled.

Claims 21, 22, 23, 24, 25, 26, 27, 28,

25, 26, 27, and 28 are rejected and are the subject of appeal.

Claims 29, 30, 31, 32, 33 and 34 are canceled.

Claims 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48 and 49 are rejected and are the subject of appeal.

IV. Status of Amendments

All amendments have been entered.

V. Summary of the Claimed Subject Matter

The claimed subject matter relates to isolated antibodies or antibody fragments that bind to a protein referred to by Appellant as Ovr110. More specifically, the claims are drawn to isolated antibodies or antibody fragments that bind specifically to a protein encoded by polynucleotide sequence SEQ ID NO:1 or to a fragment of the protein encoded by SEQ ID NO:1 which is encoded by SEQ ID NO:12 or 13. Methods for binding these antibodies on a cell are also claimed.

The polynucleotide sequence of SEQ ID NO:1 and fragments SEQ ID NO:12 and 13 are set forth in the Sequence Listing. The polynucleotide sequence of SEQ ID NO:1 is inclusive of the entire open reading frame for the protein. Further, the specification teaches in Examples 1 and 2 at pages 16-18 that SEQ ID NO:1 is an mRNA molecule and thus has a set 5' to 3' orientation.

Antibodies and antibody fragments against Cancer Specific Genes such as SEQ ID NO:1 and fragments thereof such as SEQ ID NO:12 and 13 as well as methods for use of these antibodies are described in detail in the specification, for example at pages 11-12 and 14-15. Teachings in Examples 1 and 2 relating to mRNA overexpression of Ovr110 (SEQ ID NO:1) provide further evidence of the utility of this Cancer Specific Gene as a diagnostic marker for gynecologic cancers.

VI. Grounds of Rejections to be Reviewed on Appeal

Whether claims 14, 21, 22, 23, 24, 25, 26, 27, 28, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48 and 49 meet the utility requirement of 35 U.S.C. 101.

Whether claims 14, 21, 22, 23, 24, 25, 26, 27, 28, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48 and 49 meet the enablement requirement of 35 U.S.C. 112, first paragraph.

Whether claims 14, 21, 22, 23, 24, 25, 26, 27, 28, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48 and 49 meet the written description requirement of 35 U.S.C. 112, first paragraph.

VII. Arguments

Claims of the instant application are drawn to isolated antibodies or antibody fragments that bind specifically to a protein encoded by a polynucleotide sequence SEQ ID NO:1, or a fragment of the protein encoded by polynucleotide sequence SEQ ID NO:1, wherein the fragment is encoded by polynucleotide sequence SEQ ID NO:12 or 13. Also claimed are methods for binding these antibodies or antibody fragments on a cell by contacting the cell with the isolated antibody or antibody fragment. The pending claims stand rejected under 35 U.S.C. 101 and 35 U.S.C. 112, first paragraph, as not being supported by either a substantial utility or a well established utility has been maintained. The pending claims also stand rejected under 35 U.S.C. 112, first paragraph for failing to meet the written description requirement. The underlying question in each rejection raised by the Examiner is whether express written disclosure in the specification of an amino acid sequence for a protein encoded by an expressly disclosed full length nucleic acid, in this case SEQ ID NO:1, is required to meet the utility and enablement requirements as set forth in 35 U.S.C. 101 and 112, first paragraph, as well as the written description requirement set forth in 35 U.S.C. 112, first paragraph, for a claim drawn to an isolated antibody or antibody fragment that binds specifically to a protein encoded by polynucleotide sequence SEQ ID NO:1.

Appellant respectfully submits that in the instant patent application, with the instant facts, express written disclosure in the specification of an amino acid sequence for a protein encoded by an expressly disclosed full length nucleic acid, in this case SEQ ID NO:1, is not required to meet the statutory requirements of 35 U.S.C. 101 and 35 U.S.C. 112, first paragraph.

A. Rejection under 35 U.S.C. 101

With respect to the rejection of the pending claims under 35 U.S.C. 101, MPEP 2107.02 and the case law are clear,

> as a matter of Patent Office practice, a specification which contains a disclosure of utility which corresponds in scope to the subject matter sought to be patented must be taken as sufficient to satisfy the utility requirement of §101 for the entire claimed subject matter unless there is reason for one skilled in the art to question the objective truth of the statement of utility or its scope.

In re Langer, 503 F.2d 1380, 1391, 183 USPQ 228, 297 (CCPA 1974)(emphasis in original).

The originally filed specification contains a disclosure of utility which corresponds in scope to the claimed subject matter. Specifically, at page 3, lines 26 through page 4, line 2, as well as page 7, lines 2-35, of the specification, it is taught that nine Cancer Specific Genes (CSGs) have been identified and refer, among other things, to native proteins expressed by the genes comprising the polynucleotide sequences of any of SEQ ID NO: 1, 2, 3, 4, 5, 6, 7, 8 or 9, the native mRNAs encoded by the genes comprising any of the polynucleotide sequences of SEQ ID NO: 1, 2, 3, 4, 5, 6, 7, 8 or 9 or the actual genes comprising any of the polynucleotide sequences of SEQ ID NO: 1, 2, 3, 4, 5, 6, 7, 8 or 9. It is also taught that fragments of the CSGs such as those depicted in SEQ ID NO:10, 11, 12, 13 or 14 can also be detected. Further, at page 6, lines 3 through 20, of the specification it is stated that:

> antibodies against CSG or fragments of such antibodies which can be used to detect or image localization of CSG in a patient for the purpose of detecting or diagnosing selected cancers. Such antibodies can be polyclonal or monoclonal, or prepared by molecular biology techniques. The term "antibody", as used herein and throughout the instant specification is also meant to include aptamers and single-stranded oligonucleotides such as those derived from an *in vitro* evolution protocol referred to as SELEX and well known to those skilled in the art. Antibodies can be labeled with a variety of detectable labels including, but not limited to, radioisotopes and paramagnetic metals. These antibodies or fragments thereof can also be used as therapeutic agents in the treatment of diseases characterized by expression of a CSG. In therapeutic applications, the antibody can be used without or with derivatization to a cytotoxic agent

> such as a radioisotope, enzyme, toxin, drug or a prodrug.

In addition, at page 11, line 5 through page 12, line 7 of the instant specification, assay techniques that can be used to determine levels of a CSG of the present invention, in a sample derived from a patient are described. Included in the assays taught in the specification are radioimmunoassays, immunohistochemistry assays, competition assays, Western Blot analyses and ELISA assays, all of which are well known to those of skill in the art and involve antibodies to the CSG. ELISA and competition assays are described in detail at page 11, line 16 through page 12, line 7 and page 12, lines 8 through 29, respectively, and each assay is explicitly stated to require an antibody specific to CSG. In vivo antibody uses are also taught in the specification at page 14, line 5 through page 15, line 27 in a subsection of the specification entitled ***"In Vivo Antibody Use"***. Therein is it stated that:

> Antibodies against CSG can also be used *in vivo* in patients suspected of suffering from a selected cancer including lung cancer or gynecologic cancers such as ovarian, breast, endometrial or uterine cancer [and that] antibodies against a CSG can be injected into a patient suspected of having a selected cancer for diagnostic and/or therapeutic purposes.

It is further stated that "use of antibodies for *in vivo* diagnosis is well known in the art" and several examples of antibodies used as in vivo diagnostics in cancer are provided as evidence to support this statement. In addition, details on administering antibodies against a CSG for the purpose of diagnosing or staging of the disease status of the patient are set forth as well as injection of an antibody against a CSG for therapeutic benefit. Again, several examples of antibodies used therapeutically in cancer are provided as evidence of the credibility of the substantial asserted utility of the instant claimed invention.

Finally, Examples 1 and 2 set forth at pages 16-18 of the specification describe experiments relating to mRNA overexpression of SEQ ID NO:1, also referred to as Ovr110, demonstrative to the skilled artisan of its utility as a diagnostic marker for gynecologic cancers.

Thus, teachings of the original, as-filed specification clearly assert a substantial utility for the claimed invention.

Further, Appellant provided with the Office Action response filed May 3, 2005 confirming evidence that the claimed invention is useful in the manner taught in the originally filed application. Appellant provided two publications confirming teachings in the originally filed specification that elevated mRNA expression of Ovr110 (SEQ

ID NO:1) in gynecologic cancer tissues as taught at page 17-24 of the specification correlates with measurable protein levels in gynecologic cancers.

Specifically, Tringler et al. published results from a study designed to investigate the expression of the DDO110 protein, also known as Ovr110 (the protein encoded by SEQ ID NO:1), which is homologous to B7-H4 (see page 1842, col. 2 of Tringler et al.), in normal breast and in primary and metastatic breast carcinoma in the March 1, 2005 issue of Clinical Cancer Research. A copy of this reference is provided herewith as Evidence Appendix A. Ovr110 protein exhibited nearly ubiquitous expression in breast cancer, independent of tumor grade or stage and is suggested to have a critical role in breast cancer biology. At page 1842, col. 2, Tringler et al. teaches that this gynecologic cancer marker was initially identified and characterized via quantitative PCR analysis (such as set forth in the instant specification at pages 17-24). Experiments and results set forth at pages 1843-1845 of Tringler et al. confirm the substantial asserted utility of an antibody of the claimed invention in detecting overexpression of Ovr110 in cancer tissues in accordance with methods such as taught at page 11-15 of the originally filed specification.

Also provided by Appellant with the Office Action response filed May 3, 2005 was a reference by Salceda et al. available publicly online on March 9, 2005 and published in the May 15, 2005 issue of Experimental Cell Research, Volume 306, number 1 at pages 128-141. A copy of this reference is provided herewith as Evidence Appendix B. In the results section at page 132 (col. 2), Salceda et al. teach that the detection of Ovr110 protein (the protein encoded by SEQ ID NO:1) (referred to therein as B7-H4 or DDO110)) "in human breast and ovarian cancers but not in most normal adult tissues by Western blot is in good agreement with mRNA expression data." Further, in the Discussion at page 139, it is taught that "B7-H4 mRNA was overexpressed in serous ovarian cancer and a majority of breast cancers with little or no expression in a variety of normal tissues surveyed" and that "Western blots with a monoclonal antibody against B7-H4 showed that B7-H4 protein expression reflected this mRNA distribution".

MPEP 2107.02B teaches where an applicant has specifically asserted that an invention has a particular utility, the assertion cannot simply be dismissed by Office personnel as being "wrong" even when there may be reason to believe that the assertion is not entirely accurate. Rather, Office personnel must determine if the assertion of utility is credible (i.e. whether the assertion of utility is believable to a person of ordinary skill in the art based on the totality of evidence and reasoning provided). An assertion is credible unless (A) the logic underlying the assertion is seriously flawed, or (B) facts upon which the assertion is based are inconsistent with the logic underlying the assertion.

For the instant invention, the logic underlying the asserted utility is clearly not flawed. Nor are the facts upon which the assertion is based inconsistent with the logic underlying the assertion. Instead, the asserted utility has been confirmed in publications by Tringler et al. and Salceda et al. Accordingly, the asserted utility of the instant invention must be credible.

The Court in *In re Rinehart*, 531 F.2d 1048, 1052, 189 USPQ 143, 147 (CCPA 1976), held that "[w]hen the record as a whole would make it more likely than not that the asserted utility for the claimed invention would be considered credible by a person of ordinary skill in the art, the Office cannot maintain the rejection. Accordingly, Appellant submitted with their Reply mailed April 22, 2008, a Declaration by Dr. Patrick Sluss. A copy of Dr. Sluss' Declaration is provided herewith as Evidence

Appendix C. Paragraphs 1 through 3 of Dr. Sluss' Declaration make clear that he is one of skill in this art. After review of the instant application and in particular data presented in Examples 1 and 2 of the patent application relating to mRNA overexpression of Ovr110, Dr. Sluss believed Ovr110 to be useful as a diagnostic marker for gynecologic cancers. See specifically paragraph 5 of Dr. Sluss' Declaration. Thus, Appellant has also provided evidence that the asserted utility for the claimed invention is considered credible by a person of ordinary skill in the art.

In contrast, Office personnel have failed to provide evidence sufficient to show that the statement of asserted utility would be considered "false" by a person of ordinary skill in the art as required by MPEP 2107.02. While several literature references were cited by Office personnel in the Office Action mailed January 3, 2005 in support of the suggestion that steady state levels of mRNA do not necessarily correlate with steady state levels of proteins, none of the references are related to the protein encoded by SEQ ID NO:1. Instead, these references report unique findings of scientific interest wherein researchers unexpectedly found that protein and mRNA levels did not always correlate for a unique group of proteins. Copies of the references cited in the January 3, 2005 Office Action are provided herewith for convenience in Evidence Appendix D. However, these references are not representative of the art for proteins in general wherein mRNA levels correlate quite well with protein levels.

Appellant has shown multiple places in the originally filed specification wherein a specific utility for the claimed invention is asserted. Appellant has provided confirming evidence via published references of this utility. Finally, Appellant submitted a Declaration by one skilled in the art stating that the data of Examples 1 and 2 of the patent application demonstrates "the utility of Ovr110 as a diagnostic marker for gynecologic cancers." Accordingly, the evidence as a whole, makes clear that the asserted utility for the claimed invention is credible and further maintenance of any rejection under 35 U.S.C. 101 is improper. See *In re Rinehart*, 531 F.2d 1048, 1052, 189 USPQ 143, 147 (CCPA 1976).

B. Enablement Rejection under 35 U.S.C. 112, first paragraph

With respect to the rejection of the pending claims under 35 U.S.C. 112, for lack of enablement, the test of enablement is whether one reasonably skilled in the art could make or use the claimed invention from the disclosures in the patent coupled with information known in the art without undue experimentation. See MPEP 2164.01. Thus, the test of enablement is not whether any experimentation is necessary but whether, if experimentation is necessary it is undue. In re Angstadt, 537 F.2d 498, 504, 190 USPQ 214, 219 (CCPA 1976). If the art typically engages in such experimentation, it is not considered undue. See In re Certain Limited-Charge Cell Culture Microcarriers, 221 USPQ 1165, 1174 ((Int'l Trade Comm'n 1983), aff'd sub nom., Massachusetts Institute of Technology v. A.B. Fortia, 774 F.2d 1104, 227 USPQ 428 (Fed. Cir. 1985). Further, information well known in the art does not need to be described in detail in the specification. MPEP 2163 at page 2100-170 and Hybritech, Inc. v. Monoclonal Antibodies, Inc., 802 F.2d 1367, 1379-80, 231 USPQ 81, 90 (Fed. Cir. 1986).

Detailed guidance for the skilled artisan to make and use the claimed invention is provided in teachings throughout the specification. Teachings in Examples 1 and 2 relating to mRNA overexpression of Ovr110 (SEQ ID NO:1) are demonstrative to the skilled artisan of its utility as a diagnostic marker for gynecologic cancers. Uses for the protein encoded by SEQ ID NO:1 and antibodies against Cancer Specific Genes such as SEQ ID NO:1 are described in detail in the specification, for example at pages 11-12 and 14-15 of the instant application.

Further, Appellant submitted with their Reply filed November 22, 2005 a Declaration by inventor Dr. Susana Salceda which makes clear that any additional tools necessary to make and use the invention as claimed were well known and were used routinely by those the skilled artisan with information such as taught in the specification. As copy of Dr. Salceda's Declaration is provided herewith in Evidence Appendix E. As discussed in detail in paragraph 6 of Dr. Salceda's Declaration, protein sequences and/or open reading frames were routinely obtained by those skilled in the art at the time of filing the instant patent application based upon information such as provided in the instant specification. In particular, the specification teaches in Examples 1 and 2 at pages 16-18 that SEQ ID NO:1 is an mRNA molecule and thus has a set 5' to 3' orientation. From this information, one skilled in the art knows that the protein is encoded in the forward (5' to 3') direction of SEQ ID NO:1. This characteristic taught in the originally filed specification limits the potential frame translations to three possibilities. Further, as explained in paragraph 6 of Dr. Salceda's Declaration, one skilled in the art understands that in general the open reading frame is: the frame of SEQ ID NO:1 encoding for a methionine near the 5' end; the frame encoding many amino acids; and, the frame terminating with a stop codon. Any frame with multiple stop codons can thus be ruled out. Multiple tools were available by 1998, thus preceding the September 2, 1998 priority date of the instant application, which could be used to routinely determine the protein sequence and/or open reading frame of SEQ ID NO:1 based upon the information provided in the originally filed specification. Provided with Dr. Salceda's Declaration are examples of results from three different computer programs available to those skilled in the art as of the filing date of the instant application. These examples are also provided in Evidence Appendix E. Quite clear from these examples is the fact that for this particular nucleic acid sequence, SEQ ID NO:1, there was only one possible frame for a full length protein, frame 2, with a methionine near the 5' end, encoding a protein of over 200 amino acids in length and terminating with a stop codon. Thus, as shown by the Figures of Dr. Salceda's Declaration, the results of which are described in detail in paragraph 6 of Dr. Salceda's Declaration, using only the information disclosed in the instant specification, each of these programs was able to identify the open reading frame and protein encoded by SEQ ID NO:1. This simple step required to identify the open reading frame and protein encoded by SEQ ID NO:1 using the characteristics of SEQ ID NO:1 taught in the instant specification does not constitute undue experimentation.

As additional evidence that one of skill in the art would know there was only one possible frame for a full length protein, frame 2, in SEQ ID NO:1, Appellant provided with their Reply mailed April 26, 2006 references evidencing that, in general, the sequence flanking functional initiator codons in eukaryotic mRNA sequences is a nonrandom sequence, referred to as the Kozak consensus sequence. Also provided were multiple references evidencing that it was known that, in general, the 5'-proximal ATG serves as the initiator codon for the majority of mRNAs. Copies of these references are provided herewith in Evidence Appendix F.

The Declaration by skilled artisan Dr. Patrick Sluss (see Evidence Appendix C) serves as yet additional evidence that those skilled in the art as of 1998 typically engaged in experimentation such as required in the instant application to identify antibodies binding to the protein encoded by SEQ ID NO:1. In paragraph 7, Dr. Sluss states "once the nucleic acid sequence is specified there were several approaches available to those skilled in the art in 1998 to generate antibodies that could be used to formulate tests for circulating proteins originating from the nucleic acid sequence revealed." Further in paragraph 8, Dr. Sluss states "[t]he nucleic acid sequences contain all the information needed for one skilled in the art to predict, using software tools available in 1998, all proteins that could be coded. These protein sequences could then be used in homology searches, again using software and databases available at the time, to identify target immunogens for specific antibody generation." Clear from Dr. Sluss' Declaration is once the nucleic acid sequence of SEQ ID NO:1 had been identified as an ovary specific gene associated with ovarian cancer, obtaining antibodies to a protein encoded thereby useful in an antibody-based diagnostic method was routine.

Also evidenced by both Declarations is that well known and routine to those of skill in the art at the time of filing the instant application were methods for expressing proteins encoded by a nucleotide sequence such as SEQ ID NO:1 and generating antibodies thereto. See paragraph 8 of Dr. Salceda's Declaration and paragraph 7 of Dr. Sluss' Declaration.

Thus, evidenced by the record as a whole is that teachings of the instant specification provide adequate disclosure when coupled with information known to those skilled in the art to make and use the invention as claimed without undue experimentation. Not only is the native protein encoded by the longest open reading frame (ORF) of SEQ ID NO:1, but the ORF begins with the functional protein transcription initiator codon commonly referred to as the Kozak consensus sequence. Further, the ORF begins at the 5'-proximal ATG in SEQ ID NO:1, the initiator codon for the majority of mRNAs. Both the Kozak consensus sequence and 5'-proximal ATG are well-known characteristics of the coding sequence of nucleic acids and therefore need not be expressly outlined in the specification. The disclosed structures and features of SEQ ID NO:1, coupled with the tools available at the time of filing the instant application to identify the open reading frame in a nucleic acid sequence which are outlined in Dr. Salceda's and Dr. Sluss' Declarations, clearly provide sufficient information to enable one of skill in the art to routinely make and use the instant claimed invention without undue experimentation, thus meeting the requirements of 35 U.S.C. 112, first paragraph, with respect to the enablement.

Further maintenance of this enablement rejection is therefore improper.

C. Written Description Rejection under 35 U.S.C. 112, first paragraph

Finally, with respect to the rejection of the pending claims under 35 U.S.C. 112, for written description, it is respectfully submitted that the disclosure not only distinguishes the claimed invention from other materials but also leads one of skill in the art to a conclusion that the inventors were in possession of the claimed species. In the instant application, Appellant provided in the originally filed specification the nucleic acid sequences for polynucleotide SEQ ID NO:1, and multiple fragments thereof including SEQ ID NO:12 and 13. Further, Appellant taught in the originally filed specification that SEQ ID NO:1 is an mRNA, thus establishing its 5' to 3' orientation. Polynucleotide SEQ ID NO:1 is inclusive of the entire open reading frame for the protein encoded thereby. In addition, polynucleotide SEQ ID NO:1 includes a Kozak consensus sequence, well established in the art as a sequence flanking functional initiator codons in the majority of eukaryotic mRNA sequences. Further, this Kozak consensus sequence flanks the 5'-proximal ATG, well known in the art to serve as the initiator codon for the majority of mRNAs. Thus, the nucleic acid sequence taught for polynucleotide SEQ ID NO:1 in the originally filed specification includes the classic structural characteristics well established in the art to correlate with an open reading frame of a nucleic acid sequence encoding the protein. The instant application also includes a description of various methods for making the claimed antibodies and methods for using the antibodies. Thus, the instant specification meets both policy objectives of the written description requirement. See MPEP 2163 and In re Barker, 559 F.2d 588, 592 n.4, 194 USPQ 470, 473 n.4 (CCPA 1977) and Regents of the University of California v. Eli Lilly, 119 F.3d 1559, 1566, 43 USPQ2d 1398, 1404 (Fed. Cir. 1997), cert. denied, 523 U.S. 1089 (1998).

Further, MPEP 2163 states that "in the molecular biology arts, if an applicant disclosed an amino acid sequence, it would be unnecessary to provide an explicit disclosure of nucleic acid sequences that encoded the amino acid sequence. Since the genetic code is widely known, a disclosure of an amino acid sequence would provide sufficient information such that one would accept that an applicant was in possession of the full genus of nucleic acids encoding a given amino acid sequence, but not necessarily any particular species. Cf. *In re Bell*, 991 F.2d 781, 785, 26 USPQ2d 1529, 1532 (Fed. Cir. 1993) and *In re Baird*, 16 F.3d 380, 382, 29 USPQ2d 1550, 1552 (Fed. Cir. 1994)." This acknowledgement by the USPTO that "it would be unnecessary to provide an explicit disclosure of nucleic acid sequences that encoded the amino acid sequence" (emphasis added) is made for determining nucleic acid sequences from an amino acid sequence where it is well known that degeneracy of the genetic code will result in multiple nucleic acid sequences. Since an explicit disclosure is not required to meet written description under these facts, Appellant believes it is improper to require explicit disclosure in the instant application, where the genetic code has been acknowledged to be widely known and degeneracy plays no factor whatsoever in determining an amino sequence that is encoded by a disclosed nucleic acid sequence.

The Examiner relies upon Fiers v. Revel, 25 USPQ2d 1601 and Amgen v. Chugai Pharmaceutical Co. Ltd. 18 USPQ 1016 to suggest that adequate written description requires more than a mere statement that it is part of the invention and reference to a potential method of isolation; the compound itself it required. However, the facts in those cases are very different to those herein. In those cases, the claims were drawn to nucleic acid sequences for which no sequence information whatsoever was set forth in the patent application.

More relevant to the instant application are more recent decisions from the Court of Appeals for the Federal Circuit such as *Falkner v. Inglis*, 448 F.3d 1357, 1366, 79 USPQ2d 1001, 1007 (Fed. Cir. 2006) wherein the Federal Circuit explained that, "(1) examples are not necessary to support the adequacy of a written description; (2) the written description standard may be met ... even where actual reduction to practice of an invention is absent; and (3) there is no per se rule that an adequate written description of an invention that involves a biological macromolecule must contain a recitation of known structure" and *Capon v. Eshhar*, 418 F.3d 1349, 1357, 76 USPQ2d 1078, 1085 (Fed. Cir. 2005) wherein the Court state that "The 'written description' requirement must be applied in the context of the particular invention and the state of the knowledge... As each field evolves, the balance also evolves between what is known and what is added by each inventive contribution." Also see MPEP 2163.

The genetic code, coupled with reasonable predictability associated with a proximal ATG and Kozak sequences, establishes a strong structural correlation between a nucleic acid sequence and a protein encoded thereby. A strong structural correlation between an encoded protein and antibodies raised thereto is also well-established and has been recognized by the Courts in *Hybritech, Inc. v. Monoclonal Antibodies, Inc.*, 802 F.2d 1367, 1384, 231 USPQ 81, 94 (Fed. Cir. 1986) , *cert. denied*, 480 U.S. 947 (1987), *cert. denied*, 480 U.S. 947 (1987). Further, Appellant discloses in the specification that antibodies raised against a protein encoded by the disclosed nucleic acid sequence, SEQ ID NO:1, are useful in detecting gynecologic cancers and lung cancer. Evidence confirming the disclosed utility has been submitted. Accordingly, one skilled in the art would be able to predict with a reasonable degree of confidence the structure of the claimed invention from a recitation of its function. Express disclosure of the structure of the protein or antibodies thereto is therefore not required in the instant application to meet the written description requirement of 35 U.S.C. 112.

Finally, MPEP 2163 states "[i]n most technologies which are mature, and wherein the knowledge and level of skill in the art is high, a written description question should not be raised for claims present in the application when originally filed, even if the specification discloses only a method of making the invention and the function of the invention. See, e.g., *In re Hayes Microcomputer Products, Inc. Patent Litigation*, 982 F.2d 1527, 1534-35, 25 USPQ2d 1241, 1246 (Fed. Cir. 1992). In 1986, the Court of Appeals for the Federal Circuit found that raising monoclonal antibodies is conventional or well known to one of ordinary skill in the art and need not be disclosed in detail. *Hybritech Inc. v. Monoclonal Antibodies, Inc.*, 802 F.2d 1367 (Fed. Cir. 1986), *cert. denied*, 480 U.S. 947 (1987). Thus, as of 1998, when the instant patent application was filed, raising antibodies was a "mature" technology. Further, antibodies and methods of their use were described in detail at pages 6, 11-12 and 14-15 of the original specification and were claimed in originally filed claims 9 through 13. Thus, a written description question should be raised with respect to the instant claimed invention. See, e.g., *In re Hayes Microcomputer Products, Inc. Patent Litigation*, 982 F.2d 1527, 1534-35, 25 USPQ2d 1241, 1246 (Fed. Cir. 1992).

Further maintenance of this written description rejection is therefore improper.

D. Conclusion

Express disclosure in the specification of the amino acid sequence of a protein or antibodies thereto should not be required in the instant application to meet the written description and enablement requirements of 35 U.S.C. 112, first paragraph or the utility requirements of 35 U.S.C. 101 with respect to the instant claimed invention.

Written description, enablement and utility are all determined with respect to a person of ordinary skill in the art. Accordingly, during the prosecution of this case, Appellants submitted Declarations by two different persons of ordinary skill in the art, specifically, Dr. Susana Salceda and Dr. Patrick Sluss addressing in detail how each understood the information disclosed in the instant specification to show possession of the claimed invention by the inventors and how each could perform experimentation routine as of the filing date of the instant application to make and use the instant claimed invention in accordance with the utility taught in the patent application.

During prosecution of this case, Appellants submitted evidence confirming the utility of the claimed invention in accordance with teachings of the specification.

Appellants also provided during prosecution of this case, evidence in the form of literature references and computer programs available prior to the filing date of the instant application demonstrative of the characteristics of the disclosed nucleic acid sequence being enabling for the claimed invention.

Finally, Appellants have identified case law and sections of the MPEP relevant to the instant fact situation supportive of express disclosure of the structure of the protein or antibodies thereto not being required in the instant fact situation.

In contrast, it is respectfully submitted that the Examiner has provided no specific evidence or case law relevant to the instant fact situation to support the suggestion that the specification, which expressly discloses the nucleic acid sequence of SEQ ID NO:1, does not provide enough information to indicate for which proteins the claimed antibodies are specific. It is also respectfully submitted that the Examiner has failed to provide any specific evidence or case law relevant to the instant fact situation to support the suggestion that without identifying for which protein the claimed antibodies are specific in the specification, the antibodies lack utility.

Accordingly, the evidence in the prosecution history, when viewed as a whole, is indicative of the instant application meeting the written description and enablement requirements of 35 U.S.C. 112, first paragraph, and the utility requirements of 35 U.S.C. 101.

Reversal of the rejections under 35 U.S.C. 101 and 35 U.S.C. 112, first paragraph, is therefore respectfully requested.

For the reasons given in this Appeal Brief, reversal of the Examiner's rejections is requested.

Respectfully submitted,
Kathleen A. Tyrrell
Kathleen A. Tyrrell
Registration No. 38,350

DATE: January 21, 2010

LICATA & TYRRELL P.C.
66 E. Main Street
Marlton, NJ 08053
856-810-1515
E-mail: ktyrrell@licataandtyrrell.com

VIII. Claims Appendix

Claims 1-13 (canceled)

Claim 14 (previously presented): An isolated antibody or antibody fragment that binds specifically to a protein encoded by polynucleotide sequence SEQ ID NO:1.

Claims 15-20 (canceled)

Claim 21 (previously presented): The isolated antibody or antibody fragment of claim 14 wherein the antibody is a monoclonal antibody.

Claim 22 (previously presented): The isolated antibody or antibody fragment of claim 14 wherein the antibody or antibody fragment is attached to a reagent selected from the group consisting of radioactive reagents, fluorescent reagents and enzymatic reagents.

Claim 23 (previously presented): The isolated antibody or antibody fragment of claim 22 wherein the enzymatic reagent is horseradish peroxidase or alkaline phosphatase.

Claim 24 (previously presented): The isolated antibody or antibody fragment of claim 14 wherein the antibody or antibody fragment specifically binds to protein in cells, tissues, tissue extracts or bodily fluids.

Claim 25 (previously presented): The isolated antibody or antibody fragment of claim 24 wherein the antibody is a monoclonal antibody.

Claim 26 (previously presented): The isolated antibody or antibody fragment of claim 24 wherein the bodily fluids are selected from the group consisting of blood, urine, saliva and bodily secretions.

Claim 27 (previously presented): The isolated antibody or antibody fragment of claim 26 wherein blood is whole blood, plasma, or serum.

Claim 28 (previously presented): A method for binding an antibody or antibody fragment to a protein encoded by polynucleotide sequence SEQ ID NO:1 on a cell comprising contacting the cell with an isolated antibody or antibody fragment that binds specifically to a protein encoded by polynucleotide sequence SEQ ID NO:1.

Claims 29-34 (canceled)

Claim 35 (previously presented): The method of claim 28 wherein the antibody is a monoclonal antibody.

Claim 36 (previously presented): The method of claim 28 wherein the antibody or antibody fragment is attached to a reagent selected from the group consisting of radioactive reagents, fluorescent reagents and enzymatic reagents.

Claim 37 (previously presented): The method of claim 36 wherein the enzymatic reagent is horseradish peroxidase or alkaline phosphatase.

Claim 38 (previously presented): An isolated antibody or antibody fragment which binds specifically to a fragment of a protein encoded by polynucleotide sequence SEQ ID NO:1, wherein the fragment of protein encoded by polynucleotide sequence SEQ ID NO:1 is encoded by polynucleotide sequence SEQ ID NO:12 or 13.

Claim 39 (previously presented): The isolated antibody or antibody fragment of claim 38 wherein the fragment of protein encoded by polynucleotide sequence SEQ ID NO:1 is encoded by polynucleotide sequence SEQ ID NO:12.

Claim 40 (previously presented): The isolated antibody or antibody fragment of claim 38 wherein the fragment of protein encoded by polynucleotide sequence SEQ ID NO:1 is encoded by polynucleotide sequence SEQ ID NO:13.

Claim 41 (previously presented): The isolated antibody or antibody fragment of claim 38 wherein the antibody is a monoclonal antibody.

Claim 42 (previously presented): The isolated antibody or antibody fragment of claim 14 wherein the antibody or antibody fragment is attached to a cytotoxic agent.

Claim 43 (previously presented): The isolated antibody or antibody fragment of claim 42 wherein the cytotoxic agent is selected from the group consisting of drugs, toxins and radionuclides.

Claim 44 (previously presented): A method for binding an antibody or antibody fragment to a protein encoded by polynucleotide sequence SEQ ID NO:1 on a cell comprising contacting the cell with an isolated antibody or antibody fragment that binds specifically to a fragment of protein encoded by polynucleotide sequence SEQ ID NO:1, wherein the fragment of protein encoded by polynucleotide sequence SEQ ID NO:1 is encoded by polynucleotide sequence SEQ ID NO:12 or 13.

Claim 45 (previously presented): The method of claim 44 wherein the fragment of protein encoded by polynucleotide sequence SEQ ID NO:1 is encoded by polynucleotide sequence SEQ ID NO:12.

Claim 46 (previously presented): The method of claim 44 wherein the fragment of protein encoded by polynucleotide sequence SEQ ID NO:1 is encoded by polynucleotide sequence SEQ ID NO:13.

Claim 47 (previously presented): The method of claim 44 wherein the antibody is a monoclonal antibody.

Claim 48 (previously presented): The method of claim 28 wherein the isolated antibody or antibody fragment is attached to a cytotoxic agent.

Claim 49 (previously presented): The method of claim 48 wherein the cytotoxic agent is selected from the group consisting of drugs, toxins and radionuclides.

IX. Evidence Appendix

Appendix A

E E

B7-H4 Is Highly Expressed in Ductal and Lobular Breast Cancer

Barbara Tringler,[1] Shaoqiu Zhuo,[2] Glenn Pilkington,[2] Kathleen C. Torkko,[1] Meenakshi Singh,[1] M. Scott Lucia,[1] David E. Heinz,[1] Jackie Papkoff,[2] and Kenneth R. Shroyer[1]

[1]Department of Pathology, University of Colorado Health Sciences Center, Denver, Colorado and [2]diaDexus, Inc., South San Francisco, California

ABSTRACT

Purpose: **This study was designed to investigate the expression of B7-H4 protein, a member of the B7 family that is involved in the regulation of antigen-specific immune responses, in normal breast and in primary and metastatic breast carcinomas.**

Experimental Design: **Archival formalin-fixed tissue blocks from breast cancers and normal somatic tissues were evaluated for B7-H4 expression by immunohistochemistry with manual and automated image analysis. The proportion of B7-H4-positive cells and the intensity of B7-H4 staining were compared with histologic type, grade, stage, hormone receptor status, and HER-2/*neu* status.**

Results: **B7-H4 was detected in 165 of 173 (95.4%) primary breast cancers and in 240 of 246 (97.6%) metastatic breast cancers. B7-H4 staining intensity was greater in invasive ductal carcinomas [24.61 relative units (RU)] and in invasive lobular carcinomas (15.23 RU) than in normal breast epithelium (4.30 RU, $P = 0.0003$). Increased staining intensity was associated with negative progesterone receptor status ($P = 0.014$) and history of neoadjuvant chemotherapy ($P = 0.004$), and the proportion of B7-H4-positive cells was associated with negative progesterone receptor ($P = 0.001$) and negative HER-2/*neu* ($P = 0.024$) status. However, there was no statistically significant relationship between the proportion of B7-H4-positive cells or staining intensity and grade, stage, or other clinicopathologic variables. Low levels of B7-H4 expression were also detected in epithelial cells of the female genital tract, lung, pancreas, and kidney, but B7-H4 was generally absent in most other normal somatic tissues.**

Conclusions: **The nearly ubiquitous expression of B7-H4 in breast cancer, independent of tumor grade or stage, suggests a critical role for this protein in breast cancer biology.**

INTRODUCTION

Numerous therapeutic modalities are available for the adjuvant treatment of advanced breast cancer including radiotherapy, conventional chemotherapy with cytotoxic antitumor agents, hormone therapy (aromatase inhibitors, luteinizing-hormone releasing-hormone analogues), bisphosphonates, and signal-transduction inhibitors (1). The current approach to the optimal treatment selection for breast cancer is multidisciplinary and based on several factors, including clinical stage, biological characteristics of the cancer, disease recurrence, patient's age and preferences, as well as risks and benefits associated with each treatment protocol, which help clinicians to stratify patients for appropriate treatment decisions. However, despite the great variety of adjuvant treatment options, many patients either respond poorly or not at all to any of the above-described therapeutic modalities. Thus, there is a need to identify new molecular markers for breast cancer that could provide further therapeutic targets for patients that are unlikely to respond to current treatment options.

We initially identified and characterized DD-O110 as a novel gene encoding a predicted membrane glycoprotein that is overexpressed in breast and ovarian cancer with relatively little expression in normal somatic tissues, by quantitative PCR analysis of over 200 human tissue samples.[3] Based on the predicted amino acid sequence, we subsequently determined that DD-O110 is homologous to B7-H4 (also known as B7x or B7S1), a recently discovered B7 family member. B7 family members and their receptors play critical roles in the regulation of antigen-specific immune responses (2). B7-H4 ligation to its receptor BTLA on T lymphocytes results in inhibition of T-cell activation, cytokine secretion, and the development of cytotoxicity (3–6). B7-H4 mRNA but not protein expression has been detected in a wide range of normal somatic tissues, including liver, skeletal muscle, kidney, pancreas, and small bowel (3, 5). However, cell surface expression of B7-H4 protein was induced upon stimulation of T cells, B cells, monocytes, and dendritic cells in addition to a constitutive B7-H4 protein expression in lung and ovarian cancer (5). The significance of B7-H4 expression in normal or malignant nonhematopoietic cell populations has not been determined.

The present study was designed to test the hypothesis that B7-H4 protein is consistently overexpressed in primary and metastatic breast cancer and to determine if B7-H4 expression is dependent on histologic type, grade, stage, estrogen receptor (ER), progesterone receptor (PR) or HER-2/*neu* status, or with other clinical variables.

MATERIALS AND METHODS

Tissue Samples. Tissues were obtained from 173 patients with primary breast cancer who underwent surgery at the

Received 8/17/04; revised 11/11/04; accepted 12/13/04.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Requests for reprints: Kenneth R. Shroyer, Department of Pathology, University of Colorado Health Sciences Center, 4200 East Ninth Avenue Denver, CO 80262. Phone: 303-724-3060; Fax: 303-724-3712; E-mail: Ken.Shroyer@UCHSC.edu.

[3] Papkoff et al., submitted for publication.

University of Colorado Hospital, Denver, CO. Tissue blocks were assembled from the archival collections from the Department of Pathology and included 155 invasive ductal carcinomas (152 cases of ductal carcinoma of the usual type and 3 cases of invasive tubular carcinoma) and 18 lobular carcinomas of the breast. Cases with mixed patterns of histologic differentiation were excluded from the analysis. The mean age of patients at the time of diagnosis was 55.5 years (±12.8; range, 29-89 years). The tumors were classified as American Joint Committee on Cancer pathologic stage I (90 cases), stage IIa (35 cases), stage IIb (20 cases), stage IIIa (16 cases), stage IIIb (5 cases), stage IIIc (5 cases), and stage IV (2 cases). The study also included 246 breast cancer–positive lymph nodes from a subset of 27 patients who were part of the primary study population. We also evaluated normal breast tissue from women ($n = 15$) undergoing reduction mammoplasty but with no history of breast cancer. In addition, a broad spectrum of normal adult and fetal somatic tissues ($n = 314$) was evaluated for B7-H4 expression to confirm the specificity of the B7-H4 protein to breast cancer cells (Table 3).

Information on ER, PR, and HER-2/*neu* status was collected from the original surgical pathology reports. HER-2/*neu* status was determined by fluorescence *in situ* hybridization analysis (ACIS; Chromavision, San Juan Capistrano, CA). Patient survival data was provided by the University Hospital Tumor Registry for all patients. These data reported patients that had expired following the diagnosis of breast cancer but did not include information regarding disease recurrence or cause of death. This study was reviewed by the Colorado Multiple Institutional Review Board (Protocol 00-1094).

Development and Characterization of the A57.1 Antibody Directed Against B7-H4. Monoclonal antibody production and characterization was done at diaDexus (South San Francisco, CA). Seven to 8-week-old BALB/c mice were immunized twice weekly over a 5- to 6-week period with 10 μg of the recombinant B7-H4 protein, corresponding to the complete extracellular domain of the native protein. Lymphocytes were subsequently isolated and fused with P3x63Ag8.653 cells (7) to form a hybridoma using standard techniques. Hybridoma supernatants were screened by ELISA for reactivity against B7-H4 and for the absence of cross-reactivity with an unrelated recombinant protein. B7-H4-positive hybridomas were cloned by single-cell sorting using a Coulter EPICS Elite-ESP Flow Cytometer (Beckman-Coulter, Miami, FL). The A57.1 monoclonal antibody was selected for use in subsequent studies.

Western Blot Analysis. SKBR3, MCF-7, and RK3E cells were obtained from the American Type Culture Collection (Manassas, VA). RK3E cells were infected with a recombinant retrovirus expressing either B7-H4 or alkaline phosphatase used as a control. Twenty-five micrograms of protein extracts were separated on a precast 4% to 12% SDS polyacrylamide minigel (Nupage; Invitrogen, Carlsbad, CA) and transferred to an Immobilon-P polyvinylidene difluoride membrane (Invitrogen). The membrane was blocked for 1 hour at room temperature using 5% nonfat dry milk and incubated overnight with the A57.1 antibody (1 μg/mL). The blot was developed using a horseradish peroxidase linked goat anti-mouse immunoglobulin (Jackson ImmunoResearch Laboratories, Inc., West Grove, PA; 1:10,000) for 1 hour at room temperature and subsequently visualized using enhanced chemiluminescence reagent per manufacturer's directions (Amersham Biosciences, Piscataway, NY).

Immunohistochemical Staining. Formalin-fixed, paraffin-embedded tissue blocks were sectioned to 5 μm and mounted on charged glass slides (Superfrost Plus, Fisher Scientific, Pittsburgh, PA). Endogenous peroxidase activity was blocked with 3.0% hydrogen peroxide for 15 minutes. Antigen retrieval was done in a citrate buffer [20 mmol/L (pH 6.0)] at 120°C for 10 minutes. Staining was conducted on a DAKO autostainer (DakoCytomation, Carpinteria, CA) using an indirect avidin-biotin immunoperoxidase method (Vector Laboratories, Burlingame, CA). Sections were incubated at 25°C for 60 minutes with the A57.1 antibody (0.8 μg/mL). Negative controls were run on all sections at 0.8 μg/mL of a subclass-matched IgG1κ (BD PharMingen, San Diego, CA), generated against unrelated antigens. B7-H4 staining was visualized using 3,3′-diaminobenzidine (DakoCytomation). Specificity of B7-H4 staining was confirmed by a blocking experiment with preincubation of the A57.1 antibody with the full-length B7-H4 protein (7.80 ng/mL) at 25°C for 60 minutes, before immunohistochemical processing.

Evaluation of B7-H4 Staining. The proportion of B7-H4-positive cells for each case was scored on a scale from 0% to 100%. Results represent the average proportion of B7-H4-positive cells within the entire tumor area of a single representative tissue block (0-10% positive cells, >10-50% positive cells, >50-80% positive cells, and >80-100% positive cells). The B7-H4 stained slides were digitally scanned using a Zeiss Axioskop 50 microscope fitted to a Syncroscan imaging system (Syncroscopy, Cambridge, United Kingdom). Image manipulation and preparation was done using Adobe Photoshop 6.0 and image analysis of tumor and normal breast epithelium was done using Media Cybernetics Optimas 6.5 (Media Cybernetics, Silver Spring, MA). Median delta base 10 intensity values (derived from 256 Grayscale median pixel Luminosity) were corrected by subtraction of hematoxylin-based background staining and recorded as relative units (RU).

Statistical Analysis. The association of proportion of B7-H4-positive cases and proportion of B7-H4-positive cells with categorical clinicopathologic characteristics was assessed by the Fisher's Exact test or the χ^2 test where appropriate. The differences between median staining intensity and clinicopathologic variables were evaluated by the Wilcoxon rank sum test or the Kruskal-Wallis test where appropriate. Statistically significant univariate relationships were further evaluated by multivariate analysis. A log-rank test was used to test for differences in overall patient survival. $Ps \leq 0.05$ were considered statistically significant. Statistical analyses were done using SAS v8.1 (SAS Institute, Cary, NC).

RESULTS

Characterization of B7-H4 Antibody. Specificity of the A57.1 antibody for B7-H4 protein was confirmed by Western blot analysis (Fig. 1). The A57.1 antibody recognized a major protein form with a diffuse band at ~60 to 80 kDa as well as several minor species of lower molecular weight in RK3E B7-H4 cells

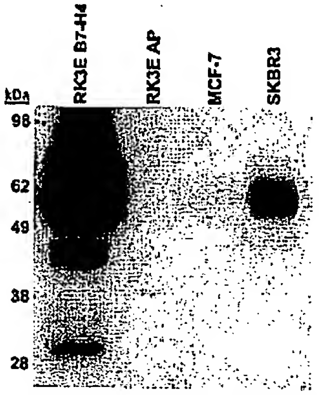


Fig. 1 Western blot analysis. The A57.1 antibody detected a major protein band at ~60 to 80 kDa and several minor bands of lower molecular weight in a RK3E cell line overexpressing B7-H4 (RK3E B7-H4). A single band of similar size was found in two breast cancer cell lines (MCF-7, SKBR3) expressing native B7-H4 mRNA but was not identified in control RK3E cells (RK3E AP).

overexpressing human B7-H4 protein, but did not detect B7-H4 protein in negative control RK3E alkaline phosphatase cells expressing alkaline phosphatase. Similar protein bands were noted in extracts of both MCF-7 (low-level B7-H4 mRNA expression) and SKBR3 (high-level B7-H4 mRNA expression) breast cancer cells. In other experiments, we have shown that the size heterogeneity observed for B7-H4 proteins in tumor tissues and cell lines is due to variable N-linked glycosylation.[4] Preincubation of the B7-H4.A57.1 antibody with the full-length recombinant B7-H4 protein completely blocked the staining in histologic sections.

Primary Breast Tumors. B7-H4 expression was detected in invasive breast cancers, including 147 of 155 (94.8%) cases of invasive ductal carcinoma and 18 of 18 (100%) cases of invasive lobular carcinoma (Table 1). In almost all cases of invasive ductal (Fig. 2*A*) and lobular (Fig. 2*B*) carcinomas, B7-H4 expression was present diffusely throughout the cytoplasm with a pronounced membranous component. However, in rare cases of ductal carcinoma ($n = 7$), the tumor cells showed only localized, incomplete cytoplasmic, and membranous staining. There was no significant association observed between B7-H4 status (positive cases versus negative cases) and any clinicopathologic variables or overall patient survival ($P = 0.910$) using the log-rank test.

The proportion of B7-H4-positive cells in most cases of ductal and lobular carcinomas was >80% of the tumor cells. However, in a small subset of B7-H4-positive carcinoma cases, <10% of the tumor cells were positive (Table 1). When the cancer cases were grouped by proportions of B7-H4-positive cells (Table 2), only PR and HER-2/*neu* were significantly associated by univariate analysis. A multivariate analysis found that grade was an effect modifier in the relationship of PR to percentage of B7-H4-positive cells. A negative PR status was a significant predictor of increasing staining intensity only in grade 3 carcinomas. HER-2/*neu* was not significant at any grade level.

The median B7-H4 staining intensity was greater in invasive ductal carcinomas (24.61 RU) and in invasive lobular carcinomas (15.23 RU) than in normal breast epithelium (4.30 RU) and the differences between these three groups were statistically significant (Table 1; Fig. 3). Univariate analysis showed that increasing B7-H4 staining intensity was associated with a negative PR status and with a history of neoadjuvant chemotherapy (azidothymidine, Adriamycin, taxotere, and cytoxan). No other statistically significant associations were observed. Multivariate analysis found that grade was again an effect modifier. A significant relationship between increasing B7-H4 staining intensity was found only in grade 3 carcinomas for those with chemotherapy. Negative PR status approached significance in grade 3 carcinomas ($P = 0.059$).

Lymph Node Metastases. B7-H4 expression was detected in tumor cells of 240 of 246 (97.6%) breast cancer–positive lymph nodes from 27 patients with nodal metastases. Within the metastatic foci, B7-H4 expression was cytoplasmic and predominantly circumferential membranous in distribution (Fig. 2*C*). The B7-H4 expression pattern of metastatic cells was always identical between individual lymph nodes from the same patient. Furthermore, B7-H4 expression in tumor cells of nodal metastases was identical to that observed in the corresponding primary tumors. Within B7-H4-negative lymph nodes with metastatic carcinoma ($n = 6$), five were from the same patient. In that patient, the primary tumor showed B7-H4 expression in only 5% of the tumor cells. The other B7-H4-negative lymph node was from a patient whose primary tumor showed B7-H4 expression in only 10% of the tumor cells. Three other lymph nodes from that same patient showed B7-H4 expression in a very low proportion of metastatic tumor cells. Focal membranous and granular cytoplasmic B7-H4 expression was also detected in scattered follicular dendritic cells of hyperplastic lymphoid follicles of lymph nodes from patients with metastatic carcinoma but was never seen in lymph nodes from patients that were negative for carcinoma.

Normal Somatic Tissue. Predominantly apical, luminal membranous B7-H4 expression was observed in ductal and lobular epithelial cells in 15 of 15 (100%) cases of normal breast tissue (Table 1; Fig. 2*D*). In one case, however, there was circumferential membranous B7-H4 expression, equivalent to that seen in breast carcinomas. B7-H4 expression was never identified in myoepithelial cells or in other cellular components of normal breast tissue.

A broad spectrum of normal adult and fetal somatic tissues was evaluated to test for the expression of B7-H4 in other cell types (Table 3). The confluent circumferential membranous pattern of expression, as seen in breast cancer cases, was never observed in normal adult somatic tissues of any anatomic site. However, apical membranous expression was noted in fallopian tubal epithelium (17 of 17), endometrial glandular epithelium (19 of 25), and occasionally in endometrial luminal surface epithelium. In addition, uniform

[4] Papkoff et al., submitted for publication.

Table 1 B7-H4 expression (no. positive cases, proportion of positive cells, and median staining intensity) in primary breast cancer and normal breast tissue

		No. cases (%) grouped by proportion of B7-H4-positive cells					Staining intensity	
Histological diagnosis	No. positive cases (%)	0-10%	>10-50%	>50-80%	>80-100%	P*	Image analysis median RU (range)	P†
Invasive ductal carcinoma‡	147/155 (94.8)	26 (17)	15 (10)	11 (7)	103 (66)	0.132	24.61 (0-75,00)	0.0003
Invasive lobular carcinoma	18/18 (100)	2 (11)	0	3 (17)	13 (72)		15.23 (0.39-55.08)	
Normal breast tissue	15/15 (100)	1 (7)	1 (7)	5 (33)	8 (53)		4.30 (1.95-13.67)	

*χ^2 test.
†Kruskal-Wallis Test.
‡Including three cases subclassified as tubular carcinoma; due to rounding, percentages in parentheses may not add up to 100%.

cytoplasmic expression without a membranous component was observed in endocervical glands (10 of 10). Focal membranous expression was detected in the bronchial epithelium of the lung (4 of 4), the columnar epithelium of the gallbladder (1 of 5), the ductal and occasionally acinar epithelium of the pancreas (10 of 10), the distal convoluted tubules of the kidney (5 of 11), and the transitional epithelium of the ureter (2 of 3) and the urinary bladder (4 of 4). Focal cytoplasmic B7-H4 expression was also noted in the pars intermedia of 1/4 sections of normal pituitary. B7-H4 cytoplasmic expression was further detected in the squamous epithelium of the larynx (2 of 3), as well as the cortex and cuticle of hair shafts and in the inner zone of the outer root sheath of hair follicles (7 of 7). All other normal somatic tissues were consistently negative for B7-H4 expression.

Within fetal tissue, B7-H4 expression was noted in the bronchial epithelium of the lung, the distal convoluted tubules and collecting ducts of the kidney, the hair follicles, the amniotic epithelium, and in cytotrophoblast cells of chorionic villi of first trimester placentas. By contrast, chorionic villi from term placentas were always negative for B7-H4 expression (Table 3).

DISCUSSION

Despite the use of a wide range of adjuvant treatment options, including radiotherapy, conventional chemotherapy with cytotoxic antitumor agents alone or in combination with endocrine therapy, bisphosphonates, and HER-2/*neu* directed therapy (trastuzumab; ref. 1), over 40,000 women will die from breast cancer in the United States in 2004 (8). Thus, new molecular targets must be defined as a first step leading to the development of novel therapeutic strategies for the treatment of breast cancer.

The current study is the first to examine the expression of B7-H4 protein in primary and metastatic breast cancer. In

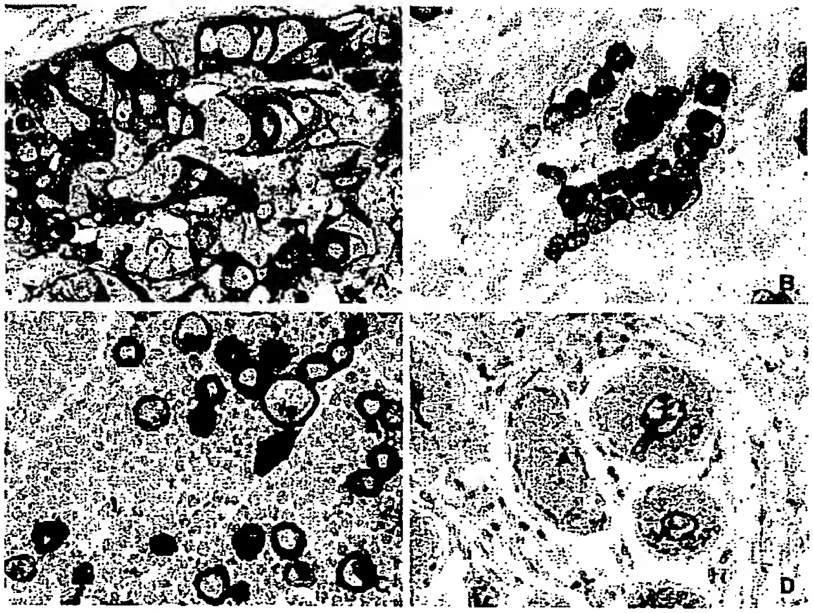


Fig. 2 Immunohistochemical detection of B7-H4 expression in breast cancer and normal breast tissue. Note strong cytoplasmic and circumferential membranous B7-H4 expression in both invasive ductal (*A*) and lobular (*B*) breast cancers. An identical pattern of B7-H4 expression is also present in metastatic breast cancer of an axillary lymph node (*C*). By contrast, predominantly apical, luminal membranous B7-H4 expression is observed in normal breast epithelium (*D*). *A*, *B*, and *C*, original magnification ×600; *D*, original magnification ×400.

Table 2 Proportion of B7-H4-positive cells and median staining intensity of 173 invasive breast cancer cases compared with clinicopathologic variables

			No. cases (%) grouped by proportion of B7-H4-positive cells					Staining intensity	
		No. cases	0-10%	>10-50%	>50-80%	>80-100%	*P**	Image analysis median RU (range)	*P*
Grade†	1	29	3 (10)	5 (17)	0	21 (72)	0.088	19.92 (0-69.92)	0.205‡
	2	56	13 (23)	7 (13)	4 (7)	32 (57)		16.60 (0-75.00)	
	3	70	10 (14)	3 (4)	7 (10)	50 (71)		26.17 (0-74.22)	
Receptor status	ER+	137	22 (18)	16 (12)	10 (7)	87 (64)	0.255	17.58 (0-75.00)	0.099§
	ER−	36	4 (11)	1 (3)	2 (6)	29 (81)		30.08 (0.78-74.22)	
	PR+	120	22 (18)	17 (14)	10 (8)	71 (59)	0.001	16.41 (0-75.00)	0.014§
	PR−	53	6 (11)	0	2 (4)	45 (85)		30.86 (0-74.22)	
	HER-2/*neu*+	25	1 (4)	2 (8)	5 (20)	17 (68)	0.024	24.61 (0-54.69)	0.924§
	HER-2/*neu*−	148	27 (18)	15 (10)	7 (5)	99 (67)		19.92 (0-75.00)	
Tumor size (cm)	≤2	108	17 (16)	15 (14)	9 (8)	67 (62)	0.063	19.92 (0-74.22)	0.120‡
	>2-5	50	11 (22)	1 (2)	2 (2)	36 (72)		16.60 (0-75.00)	
	>5	15	0	1 (7)	1 (7)	13 (87)		35.94 (8.59-63.67)	
No. lymph nodes with metastatic carcinoma	0	97	15 (15)	9 (9)	8 (8)	65 (67)	0.965	17.58 (0-69.92)	0.066‡
	1-3	31	6 (19)	2 (6)	2 (6)	21 (68)		23.44 (0-59.38)	
	>3	21	2 (10)	1 (5)	1 (5)	17 (81)		30.08 (0.78-74.22)	
	Unknown†	24	6 (25)	2 (8)	5 (21)	11 (46)		20.70 (0-66.02)	
Stage	I	90	14 (16)	13 (14)	5 (6)	58 (64)	0.082	19.92 (0-69.92)	0.194‡
	IIa	35	6 (17)	2 (6)	3 (9)	24 (69)		18.36 (0-75.00)	
	IIb	20	6 (30)	0	1 (5)	13 (65)		13.04 (0-59.38)	
	IIIa	16	0	2 (13)	0	14 (88)		36.52 (0.78-74.22)	
	IIIb+	12	2 (17)	0	3 (25)	7 (58)		16.80 (0-68.75)	
Age at diagnosis (y)	≤50	69	11 (16)	9 (13)	6 (9)	43 (62)	0.542	23.44 (0-74.22)	0.921§
	>50	104	17 (16)	8 (8)	6 (6)	73 (70)		20.90 (0-75.00)	
Neoadjuvant chemotherapy	Yes	17	1 (6)	0	2 (12)	14 (82)	0.235	41.80 (6.64-74.22)	0.004§
	No	156	27 (17)	17 (11)	10 (6)	102 (65)		19.18 (0-75.00)	

NOTE. Cases with an unknown lymph node status were excluded from the statistical analysis.

Due to rounding, percentages in parentheses may not add up to 100%.

*Fisher's exact test.

†Ductal carcinoma including three cases subclassified as tubular carcinoma.

‡Kruskal-Wallis test.

§Wilcoxon rank sum test.

addition, we evaluated B7-H4 protein expression in normal breast tissue and in a wide range of normal adult and fetal somatic tissues. B7-H4 circumferential membranous and cytoplasmic expression was observed in >95% of invasive breast cancer cases and was also detected in most nodal metastases. Univariate analysis showed a significant correlation between the proportion of B7-H4-positive cells and a negative status of PR and HER-2/*neu*. A significant association was also observed between B7-H4 staining intensity and negative PR status, and a history of treatment with neoadjuvant chemotherapy but not with other clinicopathologic variables or overall patient survival. The observed relationship between B7-H4 staining intensity level and negative PR status in primary breast cancer was not anticipated and could not be attributed to an indirect relationship with tumor grade. Thus, further studies are warranted to determine if this inverse association is due to other confounding clinicopathologic variables or could reflect a mechanistic link between B7-H4 expression and PR status.

This study provided pivotal data but not definitive evidence to support the concept that B7-H4 could be a diagnostic marker or therapeutic target for breast cancer. Both HER-2/*neu* and B7-H4 are associated with the cell surface, an important consideration for potential antibody therapeutic targets (9, 10). B7-H4 overexpression was detected in most breast carcinomas, including cases that were not candidates for hormonal or trastuzumab (Herceptin) therapy due to negative ER/PR and HER-2/*neu* status. By contrast, only weak apical cell surface expression of B7-H4 was seen in normal ductal and lobular breast epithelial cells. Focal apical membranous B7-H4 expression was also observed in the distal convoluted tubules of the kidney, ductal cells, and rare acinar cells of the pancreas, endometrial glands, and in few other normal somatic tissues. Previous studies have indicated that HER-2/*neu* is also expressed in some normal somatic tissues, including renal tubular epithelium, pancreatic acinar cells, and endometrial glands (11-14). Thus, the limited expression of B7-H4 in a subset of normal tissues does not necessarily rule out a potential role for this protein as a therapeutic target for patients with breast cancer.

Our findings that B7-H4 is not expressed in liver, small bowel, colon, and skeletal muscle are consistent with previous observations by Sica et al. (3) and Choi et al. (5). In contrast with our current study, however, Choi et al. reported that B7-H4 is not expressed in the lung, gallbladder, pancreas, kidney, ureter, urinary bladder, pituitary, or breast. The antibody used by Choi et al. was used at a dilution of 1:100 (final concentration not specified) and was noted to be reactive only with tissue frozen sections (5). By contrast, the A57.1

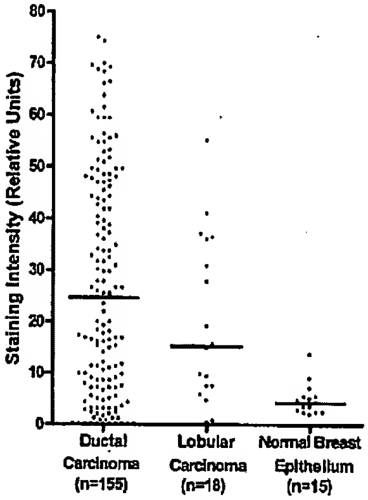


Fig. 3 Median B7-H4 staining intensity values in primary invasive ductal carcinoma, invasive lobular carcinoma, and normal breast tissue. Horizontal bars, median staining intensity within each diagnostic category.

monoclonal antibody in our study was used at a dilution of 1:2,000 (final concentration, 0.8 μg/mL) and was reactive with both frozen sections and sections from archival formalin-fixed tissue blocks. Thus, the basis for the discrepancy in the detection of B7-H4 in some normal tissues from our study compared with previous observations by Choi et al. could be due to differences in the sensitivity of the immunohistochemical staining protocols or to differences in the B7-H4 antibodies that were used.

Although the role of B7-H4 expression in malignant transformation or tumor progression has not been determined, B7 family members and their receptors are known to regulate antigen-specific immune response through inhibition of T-cell activation, cytokine secretion, and the development of cytotoxicity (2–6). Extensive laboratory and histopathologic data indicate that T-cell immune reactivity is a favorable prognostic indicator in nonmetastatic breast cancer but that the suppression of cell-mediated immunity could be critically involved in breast cancer progression (15–17). Thus, it is reasonable to hypothesize that B7-H4 overexpression could provide a mechanism for tumors to avoid detection by immune surveillance. In this light, we are currently focusing on experiments to determine if an antibody approach could inhibit tumor cell growth and/or reverse the postulated antitumor effects of B7-H4 on the immune system.

In conclusion, this study showed that B7-H4 is consistently expressed in most primary and metastatic breast carcinomas. Although B7-H4 detection was associated with negative progesterone receptor status, negative HER-2/*neu* status, and history of neoadjuvant chemotherapy, B7-H4 expression was independent of tumor grade, stage, or other clinicopathologic variables. The nearly ubiquitous expression of B7-H4 in breast carcinomas suggests that B7-H4 could be involved in breast cancer pathogenesis or tumor progression. Further studies, however, are indicated to evaluate the potential role of B7-H4 as a diagnostic marker or therapeutic target.

Table 3 B7-H4 expression (no. positive cases) in 314 normal adult and fetal somatic tissue samples

	B7-H4 positive (%)*
Normal adult tissue	
Breast (ductal and lobular cells)	16/16 (100)
Ovary	0/22 (0)
Fallopian tubal epithelium	17/17 (100)
Endometrial glands	19/25 (76)
Myometrium	0/25 (0)
Endocervical glands	10/10 (100)
Ectocervix	0/10 (0)
Thyroid	0/5 (0)
Parathyroid	0/3 (0)
Adrenal gland	0/3 (0)
Pancreas/chronic Pancreatitis	10/10 (100)
Salivary gland	0/1 (0)
Pituitary	1/4 (25)
Heart	0/5 (0)
Larynx	2/3 (67)
Lung	4/4 (100)
Esophagus	0/5 (0)
Stomach	0/5 (0)
Duodenum	0/4 (0)
Ileum	0/7 (0)
Colon/cecum	0/4 (0)
Liver	0/4 (0)
Gallbladder	1/5 (20)
Kidney (distal convoluted tubules)	5/11 (45)
Ureter	2/3 (67)
Urinary bladder mucosa	4/4 (100)
Testis	0/5 (0)
Prostate	0/5 (0)
Abdominal peritoneum	0/1 (0)
Skin	0/5 (0)
Hair follicle	7/7 (100)
Thrombus	0/4 (0)
Skeletal muscle	0/4 (0)
Synovial cyst	0/1 (0)
Bone marrow	0/5 (0)
Lymph node	0/5 (0)
Thymus	0/4 (0)
Spleen	0/5 (0)
Cerebral cortex	0/3 (0)
Cerebellum	0/3 (0)
Spinal cord	0/5 (0)
Eye	0/1 (0)
Normal fetal tissue	
Amnion	2/8 (25)
Chorion	0/8 (0)
Placental villi	2/6 (33)
Heart	0/1 (0)
Lung	1/1 (100)
Small bowel	0/1 (0)
Kidney	1/1 (100)
Skin	0/3 (0)
Hair follicle	3/3 (100)
Skeletal muscle	0/1 (0)
Cartilage	0/2 (0)
Adipose tissue	0/1 (0)

*Cases with any detectable staining (minimal focal staining or greater) were scored as B7-H4 positive.

ACKNOWLEDGMENTS

We thank Dara Hicks, Carey Arthur, David Som, and Storey Wilson for their invaluable assistance with the image analysis; the following diaDexus scientists for their contributions to the protein and antibody production: David Lowe, Shirley Vong, Yan Liu, Javier Morales, Paul Miller, Steve Lee, and Laura Corral; and Andrei Munteanu for help with the Western blots.

REFERENCES

1. Smith IE. New drugs for breast cancer. Lancet 2002;360:790–2.

2. Carreno BM, Collins M. BTLA: a new inhibitory receptor with a B7-like ligand. Trends Immunol 2003;24:524–7.

3. Sica GL, Choi IH, Zhu G, et al. B7-H4, a molecule of the B7 family, negatively regulates T cell immunity. Immunity 2003;18:849–61.

4. Prasad DVR, Richards S, Mai XM, Dong C. B7S1, a novel B7 family member that negatively regulates T cell activation. Immunity 2003;18:863–73.

5. Choi IH, Zhu G, Sica GL, et al. Genomic organization and expression analysis of B7-H4, an immune inhibitory molecule of the B7 family. J Immunol 2003;171:4650–4.

6. Wang S, Chen L. Co-signaling molecules of the B7-CD28 family in positive and negative regulation of T lymphocyte responses. Microbes Infect 2004;6:759–66.

7. Kearney JF, Radbruch A, Liesegang B, Rajewsky K. A new mouse myeloma cell line that has lost immunoglobulin expression but permit the construction of antibody-secreting hybrid cell lines. J Immunol 1979;123:1548–50.

8. American Cancer Society. Cancer facts and figures 2004. Atlanta (GA): American Cancer Society; 2004.

9. Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/*neu* oncogene. Science 1987;235:177–82.

10. Slamon DJ, Godolphin W, Jones LA, et al. Studies of the HER-2/*neu* proto-oncogene in human breast and ovarian cancer. Science 1989;244:707–12.

11. Press MF, Cordon-Cardo C, Slamon DJ. Expression of the HER-2/*neu* proto-oncogene in normal human adult and fetal tissues. Oncogene 1990;5:953–62.

12. De Potter CR, Van Daele S, Van de Vijver MJ, et al. The expression of the *neu* oncogene product in breast lesions and in normal fetal and adult human tissues. Histopathology 1989;15:351–62.

13. Zho J, Liang SX, Savas I, Banner BF. An immunostaining panel for diagnosis of malignancy in mucinous tumors of the pancreas. Arch Pathol Lab Med 2001;125:765–9.

14. Rasty G, Murray R, Lu L, Kubilis P, Benrubi F, Masood S. Expression of HER-2/*neu* oncogene in normal, hyperplastic, and malignant endometrium. Ann Clin Lab Sci 1998;28:138–43.

15. Hadden JW. The immunology and immunotherapy of breast cancer: an update. Int J Immunopathol Pharmacol 199;21:79–101.

16. Ogmundsdottir HM. Immune reaction to breast cancer: for better or for worse? Arch Immunol Ther Exp (Warsz) 2001;49 Suppl 2:S75–81.

17. Schirrmacher V, Feuerer M, Beckhove P, Ahlert, Umansky V. T cell memory, energy and immunotherapy in breast cancer. J Mammary Gland Biol Neoplasia 2002;7:201–8.

Appendix B

ELSEVIER

Available online at www.sciencedirect.com

SCIENCE @ DIRECT•

Experimental Cell Research 306 (2005) 128–141

Experimental Cell Research

www.elsevier.com/locate/yexcr

The immunomodulatory protein B7-H4 is overexpressed in breast and ovarian cancers and promotes epithelial cell transformation

Susana Salceda, Tenny Tang, Muriel Kmet, Andrei Munteanu, Malavika Ghosh, Roberto Macina, Wenhui Liu, Glenn Pilkington, Jackie Papkoff*

diaDexus, Inc., 343 Oyster Point Boulevard, South San Francisco, CA 94080, USA

Received 18 November 2004, revised version received 14 January 2005
Available online 9 March 2005

Abstract

B7-H4 protein is expressed on the surface of a variety of immune cells and functions as a negative regulator of T cell responses. We independently identified B7-H4 (DD-O110) through a genomic effort to discover genes upregulated in tumors and here we describe a new functional role for B7-H4 protein in cancer. We show that B7-H4 mRNA and protein are overexpressed in human serous ovarian cancers and breast cancers with relatively little or no expression in normal tissues. B7-H4 protein is extensively glycosylated and displayed on the surface of tumor cells and we provide the first demonstration of a direct role for B7-H4 in promoting malignant transformation of epithelial cells. Overexpression of B7-H4 in a human ovarian cancer cell line with little endogenous B7-H4 expression increased tumor formation in SCID mice. Whereas overexpression of B7-H4 protected epithelial cells from anoikis, siRNA-mediated knockdown of B7-H4 mRNA and protein expression in a breast cancer cell line increased caspase activity and apoptosis. The restricted normal tissue distribution of B7-H4, its overexpression in a majority of breast and ovarian cancers and functional activity in transformation validate this cell surface protein as a new target for therapeutic intervention. A therapeutic antibody strategy aimed at B7-H4 could offer an exciting opportunity to inhibit the growth and progression of human ovarian and breast cancers.

Keywords: B7-H4; B7x; Ovarian cancer; Breast cancer; Apoptosis

Introduction

Breast and ovarian cancer are the second and fourth leading cause, respectively, of female cancer deaths in the United States [1]. While the lifetime probability of developing breast cancer and the incidence are significantly higher than for ovarian cancer, the 5 year survival rate for breast cancer patients is notably better than for those with ovarian cancer [1]. Advances in understanding the fundamental biology and signal transduction pathways that regulate normal and malignant breast cell biology have enabled a variety of adjuvant therapeutic strategies leading to improved response rates and increased survival [2,3]. Although significant progress has been made in diagnosis and treatment of breast cancers, there is considerable opportunity for improvement and a need for additional therapeutic options [4]. Ovarian cancer is highly curable when discovered at an early stage yet a majority of these cancers progress undetected within the peritoneum resulting in diagnosis at later stages leading to a high mortality rate within a relatively short period of time [5,6]. Advances in surgery and new chemotherapeutic drugs have led to modest improvements in initial response and survival but a majority of ovarian cancer patients will relapse and die from the disease [7]. Consequently, there is an urgent need for early diagnostic tools as well as new treatment modalities.

The B7 protein family provides both stimulatory and inhibitory regulation of T cell responses, depending on which B7 ligand and receptor are engaged on the target cell [8,9]. B7-H4, also known as B7x or B7S1, is a recently discovered member of the B7 family [10–12]. B7-H4 was

* Corresponding author. Fax: +1 650 246 6598.
E-mail address: jpapkoff@diadexus.com (J. Papkoff).

0014-4827/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.yexcr.2005.01.018

shown to be a negative regulator of T cell responses in vitro by inhibiting proliferation, cell-cycle progression and cytokine production of CD4+ and CD8+ T cells [10–12]. Antigen-specific T cell responses were also impaired in mice upon treatment with a B7-H4Ig fusion protein [12]. Conversely, blockade of endogenous B7-H4 in mice using a neutralizing antibody enhanced the generation of allogeneic CTL [12]. Based on flow cytometry analysis, expression of B7-H4 protein was reported to be inducible upon stimulation of T cells, B cells, monocytes and dendritic cells whereas immunohistochemistry revealed little expression in several peripheral tissues with the exception of positive staining of some ovarian and lung cancers [12,13]. Thus, B7-H4 is postulated to attenuate inflammatory responses and perhaps serves a role in down-regulation of anti-tumor responses [10–13].

As an approach to discover and develop new diagnostic and therapeutic targets for breast and ovarian cancer, we used genomic strategies to identify genes with increased expression in human cancer compared to normal tissues. These efforts yielded a gene encoding a predicted membrane glycoprotein of unknown annotation, designated as DD-O110, whose mRNA is overexpressed in ovarian and breast cancers. DD-O110 protein is heavily glycosylated, displayed on the surface of tumor cells and overexpressed on a majority of serous ovarian carcinomas and breast cancers with little or no normal tissue expression. Based on database homology searches, we determined that DD-O110 is the same as the recently discovered immunomodulatory protein, B7-H4. Whereas B7-H4 has been implicated in regulation of immune function, our data add a new dimension to the functional role of B7-H4 in cancer by demonstrating that this protein can promote transformation and tumor formation when overexpressed in tumor epithelial cells. Together, our findings support a therapeutic antibody strategy targeting B7-H4 for cancer treatment.

Materials and methods

Discovery of DD-O110 (B7-H4), a gene upregulated in ovarian cancer

Proprietary bioinformatic algorithms were used to mine the cDNA database LifeSeq™ (Incyte Corporation, Wilmington, DE) for ESTs that were present preferentially in human ovarian cancer cDNA libraries, with low abundance in libraries from any other tissue type and disease state including normal ovary. Using these criteria, DD-O110 was discovered and full length clones were obtained from ovarian cancer cDNA using standard molecular biology methods. Comparison of the DD-O110 nucleotide sequence and predicted protein with public databases initially revealed identity to a novel protein (Genbank: gi10438801), predicted to be transmembrane, called FLJ22418. Based on routine database searches, we subsequently determined that DD-O110 was identical to the recently published B7-H4 (B7x, B7S1) [10–12].

Real-time quantitative RT-PCR (QPCR)

Tissue samples were purchased from various commercial sources including Zoion Diagnostics (Hawthorne NY) Kaplan Comprehensive Cancer Center (New York NY) and National Disease Research Exchange (Philadelphia PA). Total RNA from gross tissue was prepared using Trizol® RNA isolation reagent (Life Technologies, Grand Island, NY) and treated with RNase-free Deoxyribonuclease I (Life Technologies). For cDNA synthesis, 9 mg of RNA was added to plus-RT reaction buffer containing 50 mM KCl, 10 mM Tris–HCl pH 8.3, 5.5 mM $MgCl_2$, 200 mM of each deoxynucleotide triphosphate, 2.5 mM random hexamers, 0.4 U/ml RNase inhibitor, 1.3 U/ml MuLV reverse transcriptase and DEPC treated water (Ambion, Austin, TX) in a final volume of 500 ml. Reverse transcription reagents were obtained from Perkin-Elmer (Foster City, CA). A scaled negative control reverse transcription was set-up using 1 mg of RNA with the same reagents except without MuLV reverse transcriptase. Reverse transcription was performed in a GeneAmp, 9700 cycler (PE Applied Biosystems, Foster City, CA) with a 10 min incubation at 25°C, 1 cycle of reverse transcription at 48°C for 30 min and enzyme inactivation at 95°C for 5 min. Following reverse transcription, the RNA/cDNA mixture from the plus-RT and minus-RT reaction volumes was brought down to 1 ng/ml with Tris–EDTA buffer pH 7.0 (BioWhittaker, Walkersville, MD).

QPCR was carried out on an ABI Prism®7700 Sequence detection system (PE Applied Biosystems, Foster City, CA) in MicroAmp® optical 96-well plates, using the TaqMan® Universal PCR master mix according to the manufacturer's directions. The amplification used 10 ng of template with primers and probes designed for B7-H4 (forward primer 5′CACCAGGATAACATCTCTCAGTGAA3′, reverse primer 5′TGGCTTGCAGGGTAGAATGA3′, and probe 5′AAGCTGAAGATAATCCCATCAGGCAT3′); and the endogenous internal control ATP synthase 6 (forward primer 5′CAGTGATTATAGGCTTTCGCTCTAA3′, reverse primer 5′CAGGGCTATTGGTTGAATGAGTA3′, and probe 5′AGCCCACTTCTTACCACAAGGCACA3′). Primers and probes were synthesized by Megabases (Evanston, IL) and OligoFactory-PE Biosystems (Foster City, CA). Expression levels are represented relative to one sample named calibrator that becomes the 1× sample, and mRNA levels in all other samples are expressed as an *n*-fold difference relative to the calibrator (ABI Prism 7700 Sequence Detection System User Bulletin #2).

Cell lines

All cell lines were purchased from the American Type Culture Collection (Manassas, VA) and grown according to supplied specifications.

siRNA oligonucleotides

A siRNA was designed based on the open reading frame of the B7-H4 mRNA using methods previously described [14,15]. A random "scrambled" siRNA was used as a negative control. We also used as an additional negative control a siRNA targeting Emerin [16] and as a positive control for knockdown of an mRNA inducing apoptosis, a siRNA targeting DAXX [17]. A BLAST search against the human genome was performed with each siRNA sequence to ensure that the siRNA was target-specific. All siRNA molecules (HPP purified grade) were chemically synthesized by Xeragon Inc. (Germantown, MD). siRNAs were dissolved in sterile buffer, heated at 90°C for 1 min and incubated at 37°C for 1 h prior to use. siRNA oligonucleotides were:

B7-H4: sense 5′-GGUGUUUUAGGCUUGGUC-C(dTdT)-3′
Emerin: sense 5′-CCGUGCUCCUGGGGCUGG-G(dTdT)-3′
Scrambled: sense 5′-UUCUCCGAACGUGUCAC-GU(dTdT)-3′
DAXX: sense 5′-GGAGUUGGAUCUCUCA-GAA(dTdT)-3′

Transfection with siRNA oligonucleotides

6 × 10^4 SKBR3 cells were seeded in 12-well plates for 18–24 h prior to transfection. A final concentration of 100 nM siRNA (except 200 nM DAXX siRNA) and 1.5 μl Oligofectamine reagent (Invitrogen) were used per well of cells for transfection according to the manufacturer's protocol. siRNAs were transfected in triplicate for all experiments. Parallel wells of cells were evaluated 72 h after transfection for mRNA levels by QPCR, protein levels by Western blot and apoptosis by two different assay systems (see below). All findings were confirmed in at least 3 independent experiments. A QuantiTech SYBR Green RT-PCR kit (Qiagen Inc.) was used for QPCR evaluation of mRNA knockdown. Between 20 and 40 ng of template, RNA was used per reaction. QPCR was performed using an ABI Prism®7700 Sequence detection system as above.

Apoptosis assays

Two different assay kits were used to evaluate apoptosis. With the "Apo-ONE Homogeneous Caspase-3/7 Assay" kit (Promega Inc., Madison, WI), cells were solubilized directly on the culture plate and caspase activity, reflected by a fluorescent readout, was measured according to supplier's instructions. With the "Guava Nexin V-PE Kit" (Guava Technologies Inc.), cells were harvested by trypsinization, washed and approximately 10^5 cells were resuspended in 40 μl provided buffer and 5 μl each Annexin V (+) and 7-AAD (−) added. Following 20 min incubation on ice, cells were analyzed using the Guava PCA Flowcytometer according to manufacturer's instructions. For the anoikis assay, RK3E cells were trypsinized and resuspended in FBS-free media at 200,000 cells/ml. 1 ml aliquots were plated per well of a 12-well plate coated with poly-HEMA (Sigma-Aldrich) and incubated at 37°C for 24 h. Cells were collected and evaluated using the Guava-Nexin V-PE kit.

SDS–PAGE and Western blot analysis

72 h after transfection with siRNA, cell extracts were prepared on ice using solubilization buffer (1% NP40, 10 mM Na_2PO_4, 0.15 M NaCl) plus a complete protease inhibitor cocktail (Roche Inc.). Extracts from virus-infected or untransfected cells were similarly prepared. Pathology-verified, snap-frozen, minced tumor tissue from serous ovarian adenocarcinoma, normal ovary, breast ductal adenocarcinoma and corresponding normal adjacent tissue (Ardais Corporation Lexington MA) or normal adult tissues (ovary, spleen, bladder, kidney, liver, heart from Zoion Diagnostics) were homogenized in extraction buffer (50 mM Tris–HCl, pH 7.5, 150 mM NaCl, 1 mM EDTA, 1.0% NP40 and 0.25% DOC 0.5% plus complete protease inhibitors) followed by sonication and centrifugation to clarify the extracts. Protein extracts from human fetal normal tissues were purchased from Biochain Inc. (Hayward CA). Between 20 and 50 μg of protein extract was used per gel lane; protein equivalent concentrations were evaluated for protein level comparisons on the same gel. Pre-cast 4–12% SDS–polyacrylamide minigels with MES running buffer (Nupage; Invitrogen) were used. Gels were transferred to Immobilon-P PVDF membranes (0.45 μm pore size, Invitrogen) using 1× Nupage transfer buffer plus 10% Methanol. Membranes were blocked using 5% nonfat dry milk in PBS with 0.05% Tween-20 (PBSMT) and incubated with primary antibody overnight in PBSMT. A mouse monoclonal antibody directed against B7-H4, A57.1, was produced in-house using recombinant B7-H4 protein [18] and was used at a final concentration of 1 μg/ml. A mouse monoclonal antibody against GAPDH (Chemicon Inc.) was used at a final concentration of 2 μg/ml. Following primary antibody incubation, membranes were washed in PBS with 0.05% Tween-20 (PBST) and incubated with horseradish peroxidase linked goat anti-mouse immunoglobulin (Jackson Lab Inc.) at a 1:10,000 dilution in PBSMT. Membranes were washed with PBST followed by detection using enhanced chemiluminescence (ECL) reagent per manufacturer's directions (Amersham).

Enzymatic deglycosylation

Deglycosylation experiments were performed on protein extracts from MCF7 cells and two human serous ovarian tumor tissues using Peptide *N*-Glycosidase F (New England Biolabs, Inc, Beverly, MA) as per the manufacturer's directions. Treated samples along with untreated control samples handled in parallel were then analyzed by Western blot.

Cell surface biotinylation

Cell monolayers were washed with cold PBS and incubated on ice for 30 min with a final concentration of 0.5 μg/ml Sulfo-NHS-SS-Biotin (Pierce) in PBS. Cells were washed several times with PBS plus 25 mM Tris and then with PBS followed by extraction with solubilization buffer. Clarified supernatants were immunoprecipitated with streptavidin agarose (Pierce) followed by Western blot analysis.

Immunofluorescence

Cells, seeded on 18 × 18 mm glass coverslips, were fixed with 3.7% formaldehyde in PBS without permeabilization and then incubated for 30 min room temperature with the A57.1 antibody at a final concentration of 10 μg/ml. Cells were next washed and incubated with a secondary Cy3-labeled donkey anti-mouse (Jackson Immunoresearch Laboratories, West Grove, PA) at a concentration of 10 μg/ml for 30 min. After washing, cells were mounted in a medium containing DAPI (Vectastain, Vector, Burlingame, CA) and observed using a Zeiss fluorescence microscope Axioskop2 equipped with appropriate fluorescent filters. Micrographs were obtained with an Axiocam camera.

Immunohistochemistry

Human breast and ovarian tumors and corresponding normal tissues were obtained from National Disease Research Exchange (Philadelphia, PA). The human tissues or tumor xenograft samples were fixed in 10% neutral buffered formalin for 24 h then embedded in paraffin. Six-micrometer-thick sections were baked at 50°C, deparaffinized in Histo-clear (HS-200, National Diagnostics, Atlanta, GA) and rehydrated through decreasing ethanol concentrations into PBS. Antigen unmasking was performed by Heat Induced Epitope Retrieval (HIER) in a decloaking chamber (Biocare, Walnut Creek, CA) for 10 min in 20 mM sodium citrate buffer (pH 6.0) at 120°C, 15-17 PSI. Endogenous peroxidase activity was quenched with 3% hydrogen peroxide solution for 15 min. Slides were stained with a final concentration of 1 μg/ml A57.1 antibody using PowerVision™ IHC Homo Kit (ImmunoVision Technologies Co., Brisbane, CA) as directed. Staining was visualized by incubation with 3,3′-diaminobenzidine chromagen for 2–5 min and counterstaining with hematoxylin followed by dehydration and mounting in permount medium (Micro Mount 2000™, American Master*Tech Scientific, Inc. Lodi, CA).

Expression vector construction

B7-H4 cDNA was sub-cloned into the pLXSN retrovirus vector (BD Bioscience/Clontech) and sequence verified. The MLV LTR promotes expression of B7-H4 cDNA and an SV40 promoter drives expression of a Neo gene for G418 resistance. pLAPSN, a retroviral vector encoding alkaline phosphatase (AP), was purchased from BD Bioscience/Clontech (pLXSN-AP).

Virus production

Ecotropic virus was used to infect RK3E cells and amphotropic virus for SKOV3 cells. The pVpack-Eco plasmid (Stratagene) and pVpack-Ampho plasmid (Stratagene) were used for ecotropic and amphotropic virus packaging, respectively. 293T cells were seeded at 8 × 10^5 cells per well of a 6 well dish onto Biocoat collagen coated plates (BD). Twenty-four hours later, cells were transfected with plasmids using Lipofectamine with PLUS reagent (Invitrogen) according to the manufacturer's recommendations. Retroviral vectors, pLXSN-B7-H4 or pLXSN-AP plus pVpack-Eco/Ampho and pVpackGP (Stratagene), were transfected for 3 h after which the cells were grown overnight in DMEM containing 20% FBS. The medium was changed to DMEM with 10% FBS and virus-containing media were harvested 24 h later and filtered through a 0.45 μm polysulfonic filter.

Virus infection and selection

A final concentration of 4 μg/ml polybrene (Hexadimethrine Bromide; Sigma, St. Louis, MO) was added to fresh virus-containing medium. Cells plated the day before at a density of 3 × 10^5 cells per 100 mm^2 dish were washed with phosphate-buffered saline including Ca2$^+$ and Mg^{2+} (Cellgro). Virus containing medium was applied to the cells and incubated for 3 h at 37°C. The medium was replaced by fresh growth medium and the cells incubated at 37°C for 48–72 h at which point a final concentration of 350 μg/ml of G418 sulfate (Cellgro) was added. Following G418 selection, pools of cells were used for all subsequent experiments. Expression of ectopic proteins in the virus-infected, selected cells was verified by Western blot and expression of AP was monitored by staining.

Tumor xenograft experiments

G418-selected pools of SKOV3 cells infected with either a retrovirus expressing B7-H4 or a control retrovirus were injected subcutaneously into SCID/Beige mice (Charles River Laboratories). Ten mice were used per group and 10^7 cells in 100 μl PBS were implanted with matrigel. 100% of the mice injected with tumor cells developed tumors and tumor formation was monitored by palpation and caliper measurement. Tumor volume was calculated using the formula: (length × width^2) / 2. Data are expressed as mean group tumor volume over time. All animal experiments were performed in complete compliance with institutional guidelines.

Statistical analysis of tumor xenograft data

A single factor ANOVA was performed to test whether on the last day of measurement the tumor volumes between control and B7-H4 groups differed. The results indicated a >99.0% probability that the two groups do not have the same tumor volume. Furthermore, pairwise two-sample *t* tests assuming unequal variances with Bonferroni correction analysis were performed comparing the SKOV3-B7-H4 tumors to the SKOV3-control tumors. Analysis of data from the last day of measurement revealed that the SKOV3-B7-H4 tumors had significantly larger volumes than SKOV3-control tumors at a 99.0% confidence level.

Results

Discovery of DD-O110 (B7-H4), a gene upregulated in cancer

To identify genes that are upregulated in ovarian cancer with restricted normal tissue expression, we used bioinformatic algorithms to search for Expressed Sequence Tags (ESTs) that were present preferentially in ovarian cancer cDNA libraries, with low abundance in cDNA libraries from any other disease state or tissue type including normal ovary. Based on these criteria, one sequence identified through this comprehensive mining approach was named DD-O110. DD-O110 was also identified in a breast cancer cDNA subtraction experiment. Cloning and sequencing of a DD-O110 cDNA revealed that it encoded a predicted novel type I membrane glycoprotein with two immunoglobulin domains (Fig. 1a). Upon completion of our characterization and validation studies, we determined that DD-O110 is the same as the recently discovered negative regulatory member of the B7 family, B7-H4 (B7x, B7S1) [10–12]. We will henceforth refer to human DD-O110 as B7-H4 for consistency with the published literature.

B7-H4 mRNA is over-expressed in breast and ovarian cancers with minimal normal tissue expression

B7-H4 mRNA expression was evaluated using QPCR with a panel of 190 mRNA samples representing a variety of human cancer and normal tissues. 100% (15/15) of ductal breast adenocarcinomas and 100% (4/4) of lobular breast adenocarcinomas showed more than 2-fold overexpression of B7-H4 mRNA compared to a pool of 9 normal breast samples (Fig. 1b). Evaluation of 13 ovarian cancer samples of various subtypes and 13 normal ovarian tissues revealed that 88% (7/8) of the papillary serous adenocarcinomas expressed B7-H4 mRNA at least 2-fold higher than the average expression calculated from the 13 normal ovarian samples (Fig. 1c). The remaining ovarian cancer subtypes did not show elevated B7-H4 mRNA expression (Fig. 1c). Thirteen additional cancer types matched with normal adjacent tissue did not show significant expression of B7-H4 mRNA with the exception of uterine endometrial cancer where B7-H4 was overexpressed in 64% of the tumor samples (data not shown). Minimal expression of B7-H4 mRNA was detected in 12 normal tissue types including breast, colon, endometrium, kidney, liver, pancreas, prostate, small intestine, spleen, stomach, testis and uterus (Fig. 1c). In a separate experiment, these results were confirmed along with 11 additional normal tissues (adrenal, brain, cervix, esophagus, heart, lung, skeletal muscle, placenta, rectum, thymus and trachea) which also showed little or no B7-H4 mRNA expression (data not shown). Together, these data show that B7-H4 mRNA is overexpressed in breast and serous ovarian cancers with low or no expression in a variety of normal tissues or other cancer types.

Detection of B7-H4 protein in tumor cell lines and in human ovarian and breast tumor tissues

A mouse monoclonal antibody, A57.1, was generated that specifically recognizes B7-H4 protein using a variety of detection methods. Protein extracts from a set of cancer cell lines with and without B7-H4 mRNA expression (data not shown) were evaluated by Western blot with the A57.1 antibody. A strong, diffuse band in the 50–80 kDa size range was detected in the B7-H4 mRNA positive but not negative cell lines (Fig. 2A). The size of the major B7-H4 protein form varied between cell lines and a protein of approximately 28 kDa was also detected in some cells (Fig. 2A). Western blot analysis of three serous ovarian cancers compared to histologically normal ovarian tissue showed a diffuse B7-H4 protein band in the 50–80 kDa range in 2/3 cancers and 0/0 of the normal samples (Fig. 2B). Similarly, Western blot evaluation of three breast cancers compared to histologically normal adjacent tissue showed several prominent B7-H4 protein species in the 40–80 kDa range in 2/3 tumor samples (Fig. 2B). Faint B7-H4 protein bands were also detected in the third breast tumor sample as well as in the three normal adjacent tissue samples (Fig. 2B). No B7-H4 protein was detected in protein extracts of normal adult ovary, spleen, bladder, kidney, liver and heart (Fig. 2C). The detection of B7-H4 protein in human breast and ovarian cancers but not most normal adult tissues by Western blot is in good agreement with the mRNA expression data. A survey of human fetal tissues by Western blot detected B7-H4 protein in kidney and placenta but not in other tissues tested (Fig. 2D).

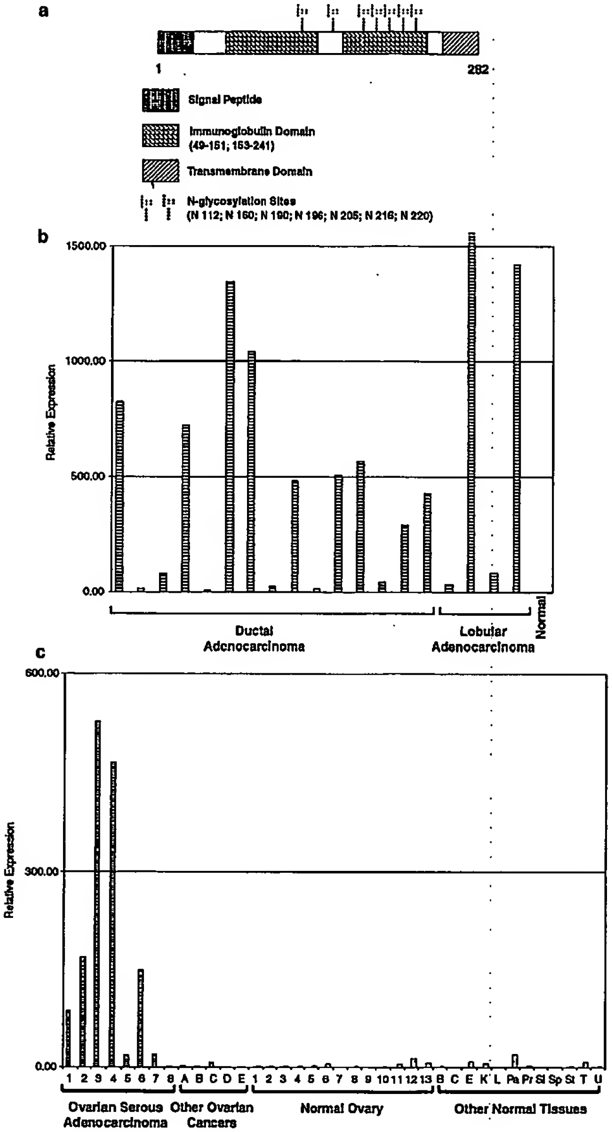


Fig. 1. (a) Schematic representation of the predicted DD-O110 (B7-H4) protein. (b) QPCR analysis of B7-H4 mRNA expression in ductal and lobular breast cancer tissues compared to pooled normal breast tissue "N." The graph shows relative mRNA expression levels in the indicated samples. (C) QPCR analysis of B7-H4 mRNA expression in ovarian cancer, normal ovary and other normal tissues. Serous ovarian cancer "1–8," mucinous and low malignant potential ovarian cancers "A–E", normal ovary "1–13" and pooled samples from normal breast "B", colon "C", endometrium "E", kidney "K", lung "L", pancreas "Pa", prostate "Pr", small intestine "SI", spleen "Sp", stomach "St", testis "T" and uterus "U".

The expression of B7-H4 was evaluated further by immunohistochemical analysis of human tissues with the A57.1 antibody. Intense circumferential membrane and cytoplasmic staining of a majority of the tumor epithelium was observed in sections of ovarian serous adenocarcimona and breast ductal adenocarcinoma (Fig. 2E, panels c,d). No B7-H4 staining was seen with normal ovarian tissue (Fig. 2E, panel a) whereas less intense staining restricted to the apical cytoplasmic membrane of ductal and lobular cells was observed in normal breast tissue (Fig. 2E, panel b). No staining of ovarian or breast tumor tissues was detected using an isotype matched control antibody (Fig. 2E, panels e, f). These immunohistochemical results are consistent with the mRNA and Western blot data described above and taken together the data show that B7-H4 mRNA and protein are over-expressed in human ovarian and breast tumor tissue compared to corresponding normal tissue.

B7-H4 is glycosylated

The protein backbone of B7-H4 is predicted to be 28.8 kDa after signal peptide cleavage, significantly smaller than the B7-H4 protein sizes observed by Western blot. The predicted B7-H4 protein sequence contains seven potential

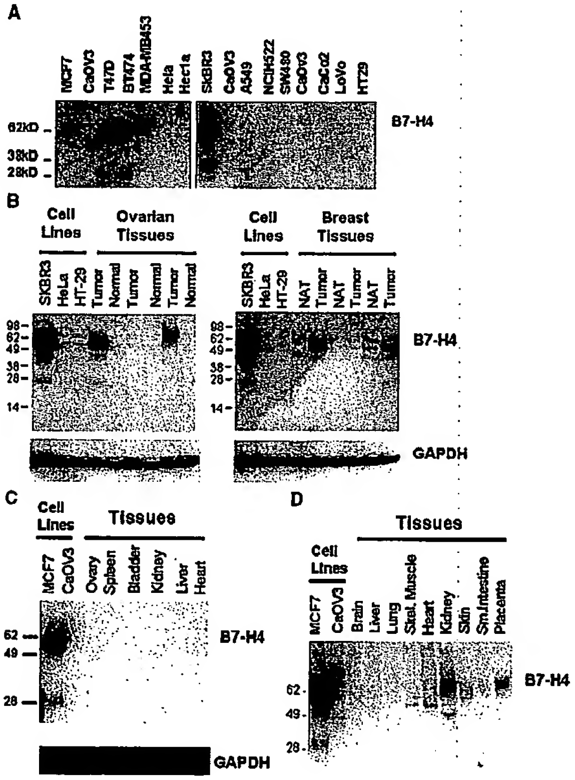


Fig. 2. Detection of B7-H4 protein in human tumor cell lines and tumor tissues by Western blot and immunohistochemistry. Protein extracts of human tumor cell lines and tissues were evaluated by SDS–PAGE followed by Western blot using the A57.1 antibody against B7-H4 and an antibody against GAPDH to verify the integrity of the tissue samples. (A) B7-H4 mRNA positive cell lines (MCF7, T47D, BT474, MDA-MB453, SKBR3) and B7-H4 mRNA negative cell lines (CaOV3, HeLa, Hec1a, A549, NCIH522, SW480, CaCo2, LoVo, HT29). (B) Serous ovarian adenocarcinomas and breast ductal adenocarcinomas with histologically normal ovarian tissue (Normal) or histologically normal adjacent breast tissue from the same patients (NAT) alongside SKBR3, HeLa and HT29 cells. (C) Adult normal tissues alongside MCF7 and CaOV3 cells. (D) Fetal normal tissues alongside MCF7 and CaOV3 cells. (E) Immunohistochemical staining of normal ovary (a), normal breast (b), ovarian serous adenocarcimona (c, e) and breast ductal adenocarcinoma (d, f) with the A57.1 antibody (a, b, c, d) or an isotype matched control antibody (e, f). The arrows indicate a normal ovarian follicle (a) and normal breast epithelium (b).

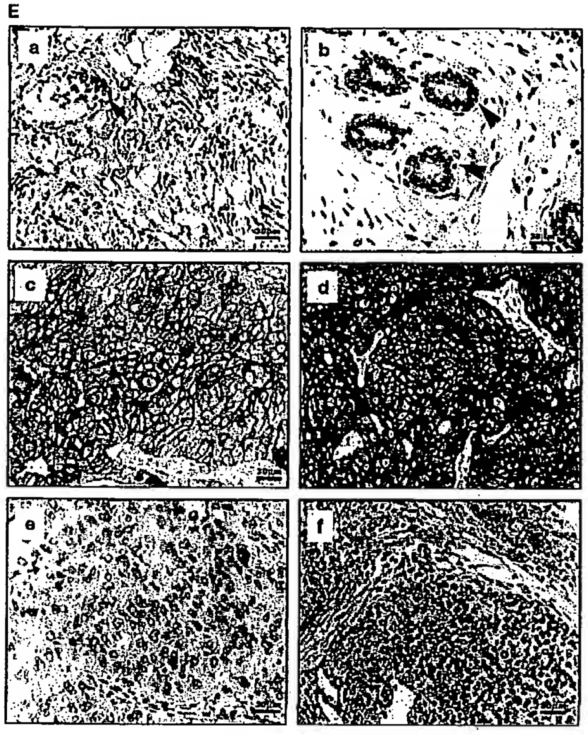


Fig. 2 (*continued*).

N-linked glycosylation sites which could account for the increased size and diffuse appearance of the protein. To test this possibility, protein extracts from MCF7 cells and two serous ovarian cancer tissues expressing B7-H4 were treated with PNGaseF, an enzyme that removes N-linked carbohydrates, followed by Western blot analysis with the A57.1 antibody. Treatment with PNGaseF reduced the B7-H4 proteins to a distinct band of approximately 28 kDa, the predicted size of the B7-H4 protein without post-translational modification (Fig. 3). These data indicate that B7-H4 is *N*-glycosylated and, based on the heterogeneous species of B7-H4 detected in different cells and tissues, suggest that the glycosylation varies between different tumors.

Identification of B7-H4 protein on the surface of breast tumor cell lines

Since B7-H4 is predicted to be a transmembrane protein, we examined whether B7-H4 could be detected on the cell surface. Live SKBR3 (B7-H4 protein positive) and HT29 (B7-H4 protein negative) cells were biotinylated to label the cell surface proteins. The cells were solubilized and biotinylated proteins were precipitated from the mixture with avidin agarose followed by Western blot analysis. B7-H4 protein was detected in the biotinylated fraction of SKBR3 but not HT29 cells thus demonstrating its cell surface localization (Fig. 4a, top panel). NaK-ATPase, a known cell surface protein, was used as a positive control and was readily detected in the biotinylated fraction of both SKBR3 and HT29 cells (Fig. 4a, middle panel). To ensure that internal cell proteins were not biotinylated, GAPDH, an abundant cytoplasmic protein, was evaluated as a negative control. As expected, GAPDH protein was not detected in the biotinylated fractions whereas it was readily detected in total cell lysates (Fig. 4a, bottom panel). Similar results were obtained upon biotinylation of MCF7 cells which also express B7-H4 protein (data not shown).

We next performed immunofluorescence to visualize the subcellular localization of B7-H4 in SKBR3 cells. Fixed, unpermeabilized SKBR3 cells as well as the B7-H4

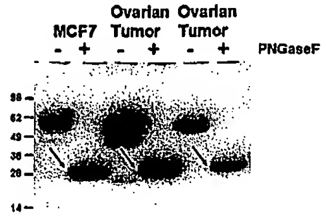


Fig. 3. B7-H4 is glycosylated. Protein extracts of MCF7 cells and two different serous ovarian cancer tissue samples were treated either with (+) or without (–) PNGaseF to remove N-linked carbohydrate. Samples were evaluated by SDS–PAGE followed by Western blot with the A57.1 antibody.

negative HT29 cells were stained with the A57.1 antibody. Specific plasma membrane labeling was observed with SKBR3 cells while the HT29 cells were negative (Fig. 4b). In agreement with these data, flow cytometry analysis using several different monoclonal antibodies against B7-H4, detected the protein on the surface of live SKBR3 cells (data not shown).

Knockdown of B7-H4 leads to increased tumor cell apoptosis

To evaluate the functional role of B7-H4 expression in tumor cells, we tested whether siRNA mediated knockdown of B7-H4 in the SKBR3 breast cancer cell line would lead to apoptosis. A B7-H4-specific siRNA diminished the level of B7-H4 mRNA in SKBR3 cells by approximately 65% at 72 h after transfection (Fig. 5a). A siRNA consisting of a scrambled sequence with no homology to any mRNAs based on blast search was used as a negative control and had no effect on B7-H4 mRNA levels (Fig. 5a). Furthermore, neither the B7-H4-specific siRNA nor the scrambled siRNA induced any non-specific knockdown of GAPDH mRNA (data not shown). Western blot analysis confirmed that the B7-H4 siRNA treated SKBR3 cells exhibited a corresponding decrease in B7-H4 protein level with no effect on GAPDH protein (Fig. 5b). The effect of the siRNAs on apoptosis was evaluated using cells treated in parallel. The B7-H4-specific siRNA led to a significant increase in apoptosis measured with an Annexin V assay whereas the scrambled control siRNA had no effect (Fig. 5c). A siRNA specific for the anti-apoptotic DAXX protein was used as a positive control for apoptosis induction since siRNA-induced knockdown of DAXX was shown previously to result in apoptosis of cultured tumor cells [17]. In SKBR3 cells, the DAXX siRNA led to approximately 65% knockdown of its corresponding mRNA (data not shown) in conjunction with increased apoptosis (Fig. 5c), similar to published findings. The extent of B7-H4 siRNA induced apoptosis in SKBR3 cells was similar to that obtained with DAXX siRNA (Fig. 5c). As an additional control to exclude any non-specific apoptotic effects of the B7-H4 siRNA, a similar experiment was performed using the HT29 tumor cell line that does not express any B7-H4 mRNA or protein. Treatment of the HT29 cells with the B7-H4 siRNA did not induce apoptosis whereas the DAXX siRNA did (data not shown).

Increased caspase activity is frequently observed during apoptosis and serves as a sensitive indicator of induced cell death [19, 20]. The activity of caspase 3 and 7 was measured in SKBR3 cells treated with either scrambled, B7-H4- or DAXX-specific siRNA. Knockdown of B7-H4 mRNA led to a significant increase in caspase activity indicating the induction of apoptosis and thus a functional role for this protein in the survival of the cells (Fig. 5d). Knockdown of DAXX mRNA also induced caspase activity as expected (Fig. 5d) [17]. QPCR performed in parallel

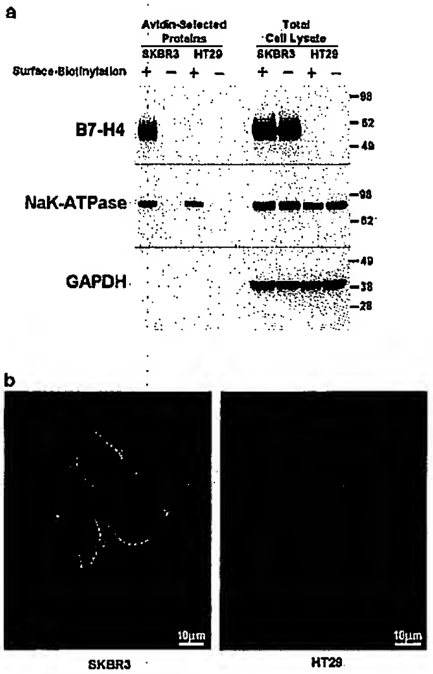


Fig. 4. Detection of B7-H4 on the cell surface. (a) Monolayers of live SKBR3 (B7-H4 positive) or HT29 (B7-H4 negative) cells were surface biotinylated (+) or mock treated (–) followed by solubilization and precipitation with avidin agarose. Avidin-affinity purified proteins were evaluated by SDS–PAGE followed by Western blot with antibodies against B7-H4 (top), NaK-ATPase (middle) or GAPDH (bottom). Protein equivalent samples of total cell lysate were evaluated in parallel. (b) Monolayers of SKBR3 (B7-H4 positive) or HT29 cells (B7-H4 negative) were used for immunofluorescence with the A57.1 antibody.

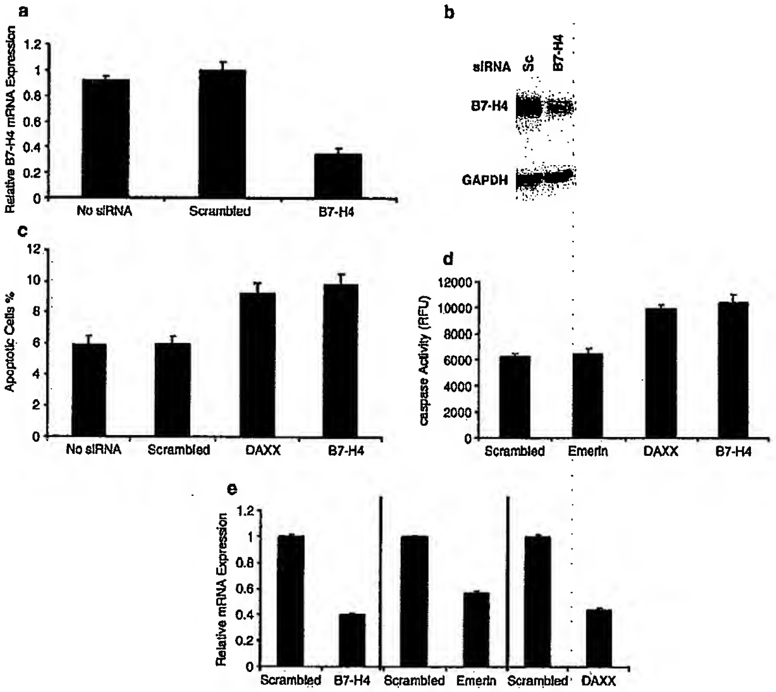


Fig. 5. Knockdown of B7-H4 mRNA and protein expression in SKBR3 cells leads to increased apoptosis. (a) SKBR3 cells were untreated (No siRNA) or treated with either scrambled- or B7-H4-specific siRNA and B7-H4 mRNA levels were evaluated by QPCR. (b) B7-H4 and GAPDH protein levels in scrambled- (Sc) or B7-H4-siRNA treated cells were evaluated by SDS–PAGE followed by Western blot with an antibody against each protein. (b) Apoptosis induction, measured by an Annexin V assay and Guava detection system, was evaluated in parallel siRNA treated cells including cells treated with siRNA specific for DAXX. (d) A different knockdown experiment was performed using scrambled, DAXX-, Emerin- or B7-H4-specific siRNAs and caspase 3/7 activity was measured. (e) B7-H4, Emerin and DAXX mRNA levels from the experiment in panel (d) were evaluated by QPCR using target-specific primers. A scrambled control was evaluated in parallel for each primer pair.

demonstrated that the B7-H4 and DAXX-specific siRNAs were able to reduce their corresponding mRNA levels by approximately 60% and 65%, respectively (Fig. 5e). In another study, knockdown of the nuclear membrane protein emerin with a specific siRNA in mammalian cells indicated that this protein is non-essential for survival [16]. SKBR3 cells were treated with siRNA specific for emerin as an additional specificity control for the effects of B7-H4 siRNA on apoptosis. While emerin siRNA led to a 50% decrease in emerin mRNA levels, there was no effect on caspase activity (Figs. 5d, e).

Overexpression of B7-H4 protects cells from apoptosis

Since knockdown of B7-H4 in tumor cells led to an increase in apoptosis, we tested conversely whether overexpression of B7-H4 could protect cells from apoptosis. A rat epithelial cell line, RK3E, was used since this model system has been employed successfully to demonstrate the effects of a variety of genes and signaling pathways important for epithelial cancers, including β-catenin, Ras and c-Myc, on apoptosis and other parameters of transformation [21–23]. B7-H4 was ectopically expressed in RK3E cells and G418-selected, polyclonal pools of retrovirus-infected cells were used for subsequent experiments to avoid clonal artifacts. Ectopic expression of alkaline phosphatase (AP) was used as a negative control. Expression of B7-H4 protein in the G418-selected cell pools was verified by Western blot (Fig. 6a). No B7-H4 protein was detected in the AP-expressing RK3E cells (Fig. 6a). Infection and selection efficiency was evaluated by staining AP-infected cell monolayers for AP activity. Essentially, all of the selected cells were positive (data not shown) and, consequently,

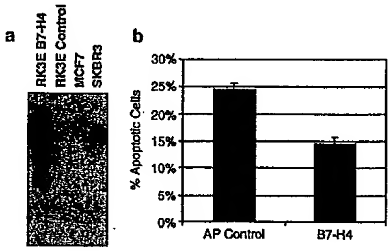


Fig. 6. Overexpression of B7-H4 protects cells from apoptosis. RK3E cells were infected with either a retrovirus expressing B7-H4 or an AP-control retrovirus followed by G418-selection. (a) Expression of B7-H4 protein in the selected cells was verified by SDS–PAGE followed by Western blot. (b) RK3E cells expressing B7-H4 or AP were evaluated using an Annexin V assay and Guava detection system to measure apoptosis in response to loss of substrate attachment after 24 h (anoikis).

most of the G418-selected cells expressed the gene of interest. AP-control and B7-H4 expressing RK3E cells were plated in suspension and the percent of cells undergoing apoptosis in response to loss of matrix adhesion (anoikis) was measured 24 h later. The cells expressing B7-H4 showed a 40% decrease in apoptosis compared to the AP-control cells (Fig. 6b) suggesting that B7-H4 overexpression can protect cells from apoptosis.

Overexpression of B7-H4 protein promotes xenograft tumor formation in SCID mice

To examine further the transforming ability of B7-H4, we evaluated the effect of ectopic B7-H4 expression on tumor formation by the human SKOV3 ovarian cancer cell line. SKOV3 cells were chosen since they represent a relevant tumor cell type for B7-H4 studies but express very low, almost undetectable, levels of endogenous B7-H4 mRNA and protein (data not shown). SKOV3 cells were infected with either a retrovirus expressing B7-H4 or a control retrovirus followed by G418-selection to enrich for infected cells. Expression of B7-H4 protein in the selected cells was verified by Western blot (Fig. 7b). Control and B7-H4 over-expressing SKOV3 cells were implanted subcutaneously into SCID/Beige mice which were then monitored for tumor formation. The SKOV3 tumor cell line is reported to form tumors as xenografts in mice [24,25]. As predicted, the control SKOV3 cells grew as xenografts where 10 out of 10 mice implanted formed tumors (Fig. 7a). All of the mice implanted with cells expressing ectopic B7-H4 protein also formed tumors, however, these tumors were larger throughout

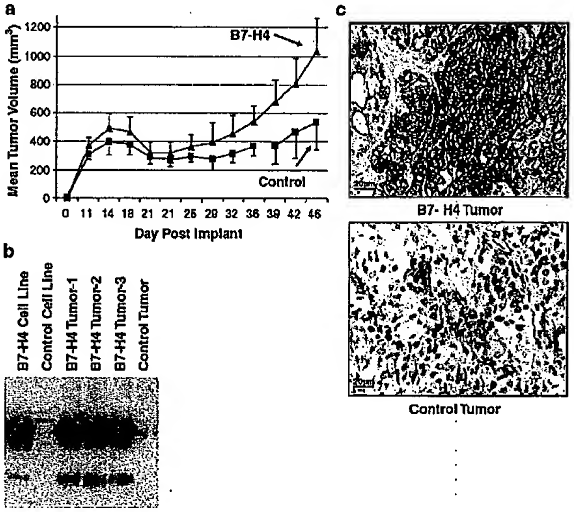


Fig. 7. Overexpression of B7-H4 facilitates xenograft tumor formation in SCID/Beige mice. SKOV3 cells were infected with either a retrovirus expressing B7-H4 or a control retrovirus followed by G418-selection. (a) Control and B7-H4-expressing SKOV3 cell pools were implanted subcutaneously into 10 SCID/Beige mice per cell type and tumor formation was monitored. The graph shows mean group tumor volume over time. (b) Expression of B7-H4 protein in the G418-selected cells used for implantation in comparison to extracts of the resulting tumor xenografts was verified by SDS–PAGE followed by Western blot. (c) Immunohistochemical staining of a B7-H4 overexpressing SKOV3 tumor and a control SKOV3 tumor with the A57.1 antibody.

the time course when compared to the control cells (Fig. 7a). Statistical analysis of the data showed that the increased size of B7-H4-SKOV3 tumors compared to control-SKOV3 tumors was significant. At the end of the study, the tumors were excised and a Western blot showed that the expression level of B7-H4 in the tumors was similar to that observed in the cells prior to implantation (Fig. 7b). Immunohistochemical staining of tumor sections with an antibody against B7-H4 also showed strong cell surface staining of a majority of the tumor cells (Fig. 7c) whereas no B7-H4 protein was detectable in the control tumors (Fig. 7c).

Discussion

We used genomic and molecular biology tools to search for genes that are upregulated in human cancers and discovered that B7-H4 mRNA was overexpressed in serous ovarian cancers and a majority of breast cancers with little or no expression in a variety of normal tissues surveyed. Western blots with a monoclonal antibody against B7-H4 showed that B7-H4 protein expression reflected this mRNA distribution. In agreement with these data, immunohistochemical staining of breast ductal adenocarcinoma and serous ovarian cancer revealed intense cell surface and cytoplasmic staining whereas no staining of normal ovary and low levels of apical staining of normal breast epithelium was detected. We also confirmed and extended these results in a comprehensive immunohistochemical analysis of human breast and ovarian cancers as well as a normal tissue panel [18,26]. In contrast to our results, another study used RT-PCR to detect B7-H4 mRNA expression in a variety of normal human tissues [12]. However, these authors indicated that no B7-H4 protein could be identified in the same tissues by immunohistochemistry. On the other hand, B7-H4 protein was shown to be inducible in T cells and antigen presenting cells and was detected by immunohistochemistry in some cases of lung cancer and ovarian adenocarcinoma [12,13]. Our findings agree with and extend these observations. Of particular note is our finding that B7-H4 overexpression is confined to the serous ovarian cancer subtype but is not observed with mucinous or low malignant potential ovarian cancers. The molecular basis for this observation is unclear but could reflect key differences in signaling pathway activation between these histologically different tumor types. We also describe for the first time the overexpression of B7-H4 mRNA and protein in a majority of breast cancers of different histological subtypes whereas significantly lower levels of B7-H4 could be detected in normal breast tissue. In other experiments, we detected overexpression of B7-H4 in uterine endometrial cancers (data not shown).

Of interest is the molecular mechanism whereby B7-H4 is apparently constitutively expressed in ovarian and breast cancer cells in contrast to the inducible expression in normal immune cells. We hypothesize that signaling pathways that are activated in tumor epithelial cells during the progression of ovarian and breast cancers could lead to the constitutive expression of B7-H4 mRNA and protein. The signals could include the participation of β-catenin, Ras or tyrosine kinase receptors, such as Her2, EGF, HGF and FGF receptors which have been implicated in development and progression of these cancers [27–29]. Experiments are in progress to uncover tumor-specific signal transduction pathways that lead to activation of B7-H4 expression.

Whereas the mouse ortholog of B7-H4 (B7S1) was reported to be a GPI-linked protein [11], we were unable to release native B7-H4 protein from a human breast cancer cell line with PI-specific PLC (data not shown) and conclude that the human protein is not GPI-linked. A similar conclusion was reached in another study based on PI-PLC experiments with transfected cells [13]. B7-H4 has a predicted C-terminal transmembrane domain and we have proven directly by cell surface biotinylation and by immunofluorescence that B7-H4 protein is localized to the surface of the human SKBR3 breast cancer cell line. Similar results were obtained with other cell types expressing B7-H4 as well as by using flow cytometry with live cells and antibodies specific for B7-H4 (data not shown). Our data also demonstrate that B7-H4 is extensively *N*-glycosylated and this glycosylation appears to be heterogeneous between tumor cell lines and tumor tissues. These differences in carbohydrate modification may reflect individual tumor-specific glyclosylation patterns and could modulate the ability of B7-H4 to interact with its receptors or potential partner proteins.

B7-H4 binds to a putative receptor expressed on activated but not naive T cells and thereby leads to inhibition of T cell activation and IL2 production [10–12]. B7-H4 does not interact with receptors for other B7 family members including CD28, ICOS, CTLA-4 or PD-1 [10] and to date the specific identity of the cell surface receptor that binds B7-H4 has not been elucidated. A receptor called B and T lymphocyte attenuator (BTLA) was proposed for B7-H4 based on indirect evidence showing that activated lymphocytes lacking BTLA exhibited reduced binding to B7-H4 compared to the wild type cells [10,30]. However, B7-H4 and BTLA do not interact with each other based on direct binding experiments ([31] and diaDexus unpublished results). It is still possible that BTLA could regulate the expression, cell surface localization or ligand binding of a B7-H4 receptor.

Despite the lack of a defined B7-H4 receptor, it is clear that B7-H4 binding to T cells leads to inhibition of T cell activation and, conversely, a neutralizing antibody against B7-H4 enhanced antigen specific T cell responses [10–12]. Since B7-H4 is overexpressed in breast and ovarian cancers, it could function to inhibit a host anti-tumor response and promote tumor escape from immune surveillance. Consequently, blockade of tumor associated B7-H4 could offer a new therapeutic opportunity for enhancement of anti-tumor immune responses. A similar function has been described for another B7 family protein, B7-H1 (PD-L1), which is normally expressed by macrophage lineage cells and is also abundant in a variety of cancers [32–36]. B7-H1 binds to its receptor, PD-1, on activated T cells and B cells leading to cell cycle arrest and apoptosis [32,37]. Tumor-expressed B7-H1 was able to inhibit T cell activation and increase apoptosis of antigen specific T cells leading to increased tumor growth in mice and these inhibitory effects could be abrogated with B7-H1 blocking antibodies [35,36,38].

The B7-H4 functional studies described here were carried out either with cultured epithelial cells in the absence of immune cell types or with human cancer cell line xenografts in SCID mice which lack most elements of a functional immune system. Therefore, our data suggesting that B7-H4 can inhibit apoptosis and promote tumor cell growth reveal a new tumor-specific function for this protein aside from its ability to modulate immune cell function. Furthermore, while B7-H4 is a membrane protein, it lacks a cytoplasmic domain and consequently it is likely to participate in signal transduction through binding to a receptor on tumor cells in either a cell-autonomous or non-autonomous fashion. B7-H4 could also function as a co-receptor for another membrane receptor with signaling functions. It will be of interest to define whether the receptor(s) and signal transduction pathways mediating the inhibitory effects of B7-H4 on T cells are same as those participating in the tumor promoting effects of B7-H4 in epithelial cancer cells. The existence of two different receptors with opposing signaling outcomes has been documented for other B7 family members, B7-H1 and B7-H2, which bind to CTLA-4 on activated T cells to mediate a negative signal and to CD28 on resting cells providing a co-stimulatory signal [8,39]. Experiments are underway to define the players and signaling pathways that are activated by B7-H4 in tumor cells.

In summary, we have shown that B7-H4 mRNA and protein are overexpressed in a majority of serous ovarian cancers and breast cancers with little or no expression in normal tissues. We show conclusively that B7-H4 protein in tumor tissue and tumor cell lines is heavily glycosylated and localized to the cell surface. Our data also provide a new function for overexpressed B7-H4 in promoting epithelial cell transformation. Together with the proposed immunomodulatory role for B7-H4, our new data validate B7-H4 as a promising new target for therapeutic antibody development and treatment of breast and ovarian cancers.

Acknowledgments

We thank Andreas Tobler for contributing to the biotinylation experiment and David Lowe, Yan Liu, Javier Morales, Laura Corral, Steve Lee and Shaoqiu Zhuo for their excellent contributions to the protein and antibody production efforts. We thank Zeanid Breyer for expert project management support and help with figures. We are very grateful to Shu-Hui Liu, Nathan Letts, Laura Corral and Thomas Mueller for critical reading of the manuscript. We thank Anne Pletcher (In-Vivo Technologies, Inc., Millbrae California) for expert advice, discussions and assistance with the mouse xenograft studies.

References

[1] American Cancer Society, Cancer statistics, 2004.

[2] V. Kaklamani, R.M. O'Regan, New targeted therapies in breast cancer, Semin. Oncol. 31 (2004) 20–25.

[3] S. Lo, S.R. Johnston, Novel systemic therapies for breast cancer, Surg. Oncol. 12 (2003) 277–287.

[4] C.A. Hudis, Current status and future directions in breast cancer therapy, Clin. Breast Cancer 4 (Suppl. 2) (2003) S70–S75.

[5] D.G. Allen, J. Coulter, Survival of patients with epithelial ovarian cancer and the effect of lymphadenectomy in those with stage 3 disease, Aust. N. Z. J. Obstet. Gynaecol. 39 (1999) 420–424.

[6] A.P. Heintz, F. Odicino, P. Maisonneuve, U. Beller, J.L. Benedet, W.T. Creasman, H.Y. Ngan, S. Pecorelli, Carcinoma of the ovary, Int. J. Gynaecol. Obstet. 83 (Suppl. 1) (2003) 135–166.

[7] J. Engel, R. Eckel, G. Schubert-Fritschle, J. Kerr, W. Kuhn, J. Diebold, R. Kimmig, J. Rehbock, D. Holzel, Moderate progress for ovarian cancer in the last 20 years: prolongation of survival, but no improvement in the cure rate, Eur. J. Cancer 38 (2002) 2435–2445.

[8] B.M. Carreno, M. Collins, The B7 family of ligands and its receptors: new pathways for costimulation and inhibition of immune responses, Annu. Rev. Immunol. 20 (2002) 29–53.

[9] S.J. Khoury, M.H. Sayegh, The roles of the new negative T cell costimulatory pathways in regulating autoimmunity, Immunity 20 (2004) 529–538.

[10] X. Zang, P. Loke, J. Kim, K. Murphy, R. Waitz, J.P. Allison, B7x: a widely expressed B7 family member that inhibits T cell activation, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 10388–10392.

[11] D.V. Prasad, S. Richards, X.M. Mai, C. Dong, B7S1, a novel B7 family member that negatively regulates T cell activation, Immunity 18 (2003) 863–873.

[12] G.L. Sica, I.H. Choi, G. Zhu, K. Tamada, S.D. Wang, H. Tamura, A.I. Chapoval, D.B. Flies, J. Bajorath, L. Chen, B7-H4, a molecule of the B7 family, negatively regulates T cell immunity, Immunity 18 (2003) 849–861.

[13] I.H. Choi, G. Zhu, G.L. Sica, S.E. Strome, J.C. Cheville, J.S. Lau, Y. Zhu, D.B. Flies, K. Tamada, L. Chen, Genomic organization and expression analysis of B7-H4, an immune inhibitory molecule of the B7 family, J. Immunol. 171 (2003) 4650–4654.

[14] S.M. Elbashir, J. Harborth, W. Lendeckel, A. Yalcin, K. Weber, T. Tuschl, Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells, Nature 411 (2001) 494–498.

[15] G.J. Hannon, RNA interference, Nature 418 (2002) 244–251.

[16] J. Harborth, S.M. Elbashir, K. Bechert, T. Tuschl, K. Weber, Identification of essential genes in cultured mammalian cells using small interfering RNAs, J. Cell Sci. 114 (2001) 4557–4565.

[17] J.S. Michaelson, P. Leder, RNAi reveals anti-apoptotic and transcriptionally repressive activities of DAXX, J. Cell Sci. 116 (2003) 345–352.

[18] B. Tringler, S. Zhuo, G. Pilkington, K. Torkko, M. Singh, M.S. Lucia, D.E. Heinz, J. Papkoff, K.R. Shroyer, B7-H4 is highly expressed in ductal and lobular breast cancer. Clinical Cancer Research, in press.

[19] N.O. Concha, S.S. Abdel-Meguid, Controlling apoptosis by inhibition of caspases, Curr. Med. Chem. 9 (2002) 713–726.

[20] F.L. Kiechle, X. Zhang, Apoptosis: biochemical aspects and clinical implications, Clin. Chim. Acta 326 (2002) 27–45.

[21] I.D. Louro, P. McKie-Bell, H. Gosnell, B.C. Brindley, R.P. Bucy, J.M.

Ruppert, The zinc finger protein GLI induces cellular sensitivity to the mTOR inhibitor rapamycin, Cell Growth Differ. 10 (1999) 503–516.

[22] Z. Weng, M. Xin, L. Pablo, D. Grueneberg, M. Hagel, G. Bain, T. Muller, J. Papkoff, Protection against anoikis and down-regulation of cadherin expression by a regulatable beta-catenin protein, J. Biol. Chem. 277 (2002) 18677–18686.

[23] T. Muller, G. Bain, X. Wang, J. Papkoff, Regulation of epithelial cell migration and tumor formation by beta-catenin signaling, Exp. Cell Res. 280 (2002) 119–133.

[24] J. Deshane, G. Cabrera, J.E. Grim, G.P. Siegal, J. Pike, R.D. Alvarez, D.T. Curiel, Targeted eradication of ovarian cancer mediated by intracellular expression of anti-erbB-2 single-chain antibody, Gynecol. Oncol. 59 (1995) 8–14.

[25] U.B. Nielsen, G.P. Adams, L.M. Weiner, J.D. Marks, Targeting of bivalent anti-ErbB2 diabody antibody fragments to tumor cells is independent of the intrinsic antibody affinity, Cancer Res. 60 (2000) 6434–6440.

[26] B. Tringler, W. Liu, L. Corrall, K. Torkko, T. Enomoto, S. Davidson, M.S. Lucia, D.E. Heinz, J. Papkoff, K.R. Shroyer, B7-H4 over-expression in ovarian cancer (Submitted for publication).

[27] B. Aunoble, R. Sanches, E. Didier, Y.J. Bignon, Major oncogenes and tumor suppressor genes involved in epithelial ovarian cancer (review), Int. J. Oncol. 16 (2000) 567–576.

[28] S.V. Nicosia, W. Bai, J.Q. Cheng, D. Coppola, P.A. Kruk, Oncogenic pathways implicated in ovarian epithelial cancer, Hematol. Oncol. Clin. North Am. 17 (2003) 927–943.

[29] R. Wu, Y. Zhai, E.R. Fearon, K.R. Cho, Diverse mechanisms of beta-catenin deregulation in ovarian endometrioid adenocarcinomas, Cancer Res. 61 (2001) 8247–8255.

[30] N. Watanabe, M. Gavrieli, J.R. Sedy, J. Yang, F. Fallarino, S.K. Loftin, M.A. Hurchla, N. Zimmerman, J. Sim, X. Zang, T.L. Murphy, J.H. Russell, J.P. Allison, K.M. Murphy, BTLA is a lymphocyte inhibitory receptor with similarities to CTLA-4 and PD-1, Nat. Immunol. 4 (2003) 670–679.

[31] J.R. Sedy, M. Gavrieli, K.G. Potter, M.A. Hurchla, R.C. Lindsley, K. Hildner, S. Scheu, K. Pfeffer, C.F. Ware, T.L. Murphy, K.M. Murphy, B and T lymphocyte attenuator regulates T cell activation through interaction with herpesvirus entry mediator, Nat. Immunol. 6 (2005) 90–98.

[32] H. Dong, S.E. Strome, D.R. Salomao, H. Tamura, F. Hirano, D.B. Flies, P.C. Roche, J. Lu, G. Zhu, K. Tamada, V.A. Lennon, E. Celis, L. Chen, Tumor-associated B7-H1 promotes T-cell apoptosis: a potential mechanism of immune evasion, Nat. Med. 8 (2002) 793–800.

[33] S.C. Liang, Y.E. Latchman, J.E. Buhlmann, M.F. Tomczak, B.H. Horwitz, G.J. Freeman, A.H. Sharpe, Regulation of PD-1, PD-L1, and PD-L2 expression during normal and autoimmune responses, Eur. J. Immunol. 33 (2003) 2706–2716.

[34] P. Loke, J.P. Allison, PD-L1 and PD-L2 are differentially regulated by Th1 and Th2 cells, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 5336–5341.

[35] S.E. Strome, H. Dong, H. Tamura, S.G. Voss, D.B. Flies, K. Tamada, D. Salomao, J. Cheville, F. Hirano, W. Lin, J.L. Kasperbauer, K.V. Ballman, L. Chen, B7-H1 blockade augments adoptive T-cell immunotherapy for squamous cell carcinoma, Cancer Res. 63 (2003) 6501–6505.

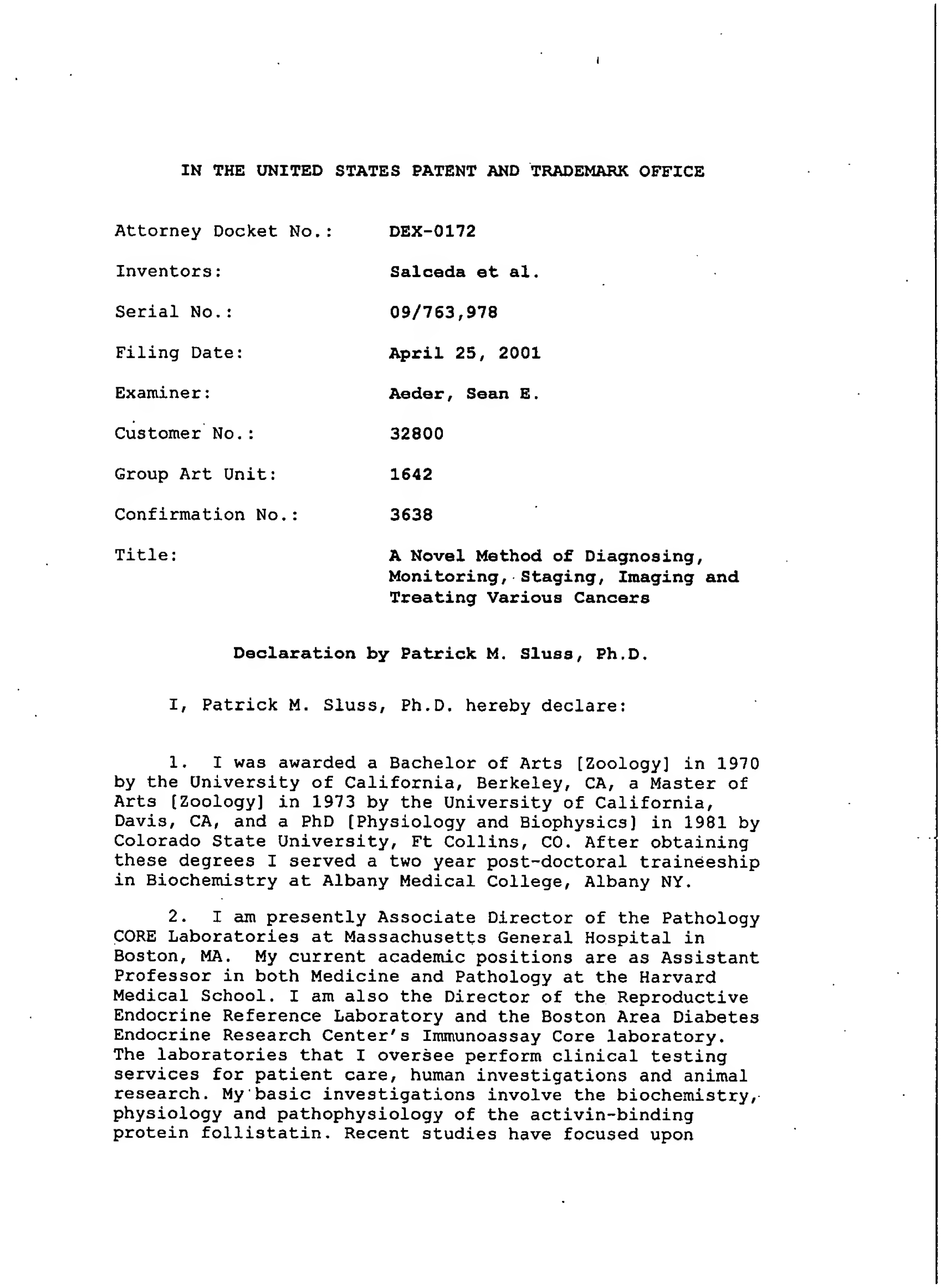
[36] S. Wintterle, B. Schreiner, M. Mitsdoerffer, D. Schneider, L. Chen, R. Meyermann, M. Weller, H. Wiendl, Expression of the B7-related molecule B7-H1 by glioma cells: a potential mechanism of immune paralysis, Cancer Res. 63 (2003) 7462–7467.

[37] H. Tamura, K. Ogata, H. Dong, L. Chen, Immunology of B7-H1 and its roles in human diseases, Int. J. Hematol. 78 (2003) 321–328.

[38] C. Blank, I. Brown, A.C. Peterson, M. Spiotto, Y. Iwai, T. Honjo, T.F. Gajewski, PD-L1/B7H-1 inhibits the effector phase of tumor rejection by T cell receptor (TCR) transgenic CD8+ T cells, Cancer Res. 64 (2004) 1140–1145.

[39] C.E. Rudd, H. Schneider, Unifying concepts in CD28, ICOS and CTLA4 co-receptor signalling, Nat. Rev., Immunol. 3 (2003) 544–556.

Appendix C



IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Attorney Docket No.:	**DEX-0172**
Inventors:	**Salceda et al.**
Serial No.:	**09/763,978**
Filing Date:	**April 25, 2001**
Examiner:	**Aeder, Sean E.**
Customer No.:	**32800**
Group Art Unit:	**1642**
Confirmation No.:	**3638**
Title:	**A Novel Method of Diagnosing, Monitoring, Staging, Imaging and Treating Various Cancers**

Declaration by Patrick M. Sluss, Ph.D.

I, Patrick M. Sluss, Ph.D. hereby declare:

1. I was awarded a Bachelor of Arts [Zoology] in 1970 by the University of California, Berkeley, CA, a Master of Arts [Zoology] in 1973 by the University of California, Davis, CA, and a PhD [Physiology and Biophysics] in 1981 by Colorado State University, Ft Collins, CO. After obtaining these degrees I served a two year post-doctoral traineeship in Biochemistry at Albany Medical College, Albany NY.

2. I am presently Associate Director of the Pathology CORE Laboratories at Massachusetts General Hospital in Boston, MA. My current academic positions are as Assistant Professor in both Medicine and Pathology at the Harvard Medical School. I am also the Director of the Reproductive Endocrine Reference Laboratory and the Boston Area Diabetes Endocrine Research Center's Immunoassay Core laboratory. The laboratories that I oversee perform clinical testing services for patient care, human investigations and animal research. My basic investigations involve the biochemistry, physiology and pathophysiology of the activin-binding protein follistatin. Recent studies have focused upon

identifying the specific epitopes directly involved with activin binding and delineating allosteric effects of activin binding in altering domain-specific antigenic epitopes in the holoprotein. In addition, my laboratory is involved in clinical studies directed toward developing methods and technological approaches for serum measurements of novel ovarian cancer markers using SELDI and other high resolution proteomic procedures. My laboratory also conducts studies of commercially available and new immunodiagnostic assays to evaluate their analytical performance and clinical utility. These studies encompass immunoassays for endocrine hormones, tumor markers, cardiac markers, and therapeutic drugs.

3. Having worked in the area of immunodiagnostics for over 20 years, I am very familiar with the methods and tools used to identify antibodies for a protein or peptide encoded by a defined nucleic acid.

4. I have reviewed the above-referenced patent application and the Office Action mailed October 22, 2007. In particular, I have reviewed the Examiner's reasoning behind his statement that "the specification does not teach the protein sequence or the open reading frame of SEQ ID NO:1" and "[t]hus . . . does not provide enough information to indicate for which protein the claimed antibodies are specific". I disagree that utility of antibodies for a diagnostic cancer marker expressed by a defined nucleic acid is dependent upon identification of "the" protein sequence or the open reading frame.

5. The Sequence Listing of the patent application sets forth a number of nucleic acid sequences including SEQ ID NO:1 and associated fragments, e.g. SEQ ID NO:10, SEQ ID NO:11, SEQ ID NO:12 and SEQ ID NO:13. A number of characteristics of SEQ ID NO:1 are described in the patent application and/or are ascertainable from the nucleic acid sequence itself. Perhaps of most importance is data presented in Examples 1 and 2 of the patent application relating to mRNA overexpression of Ovr110. This data demonstrates to me the utility of Ovr110 as a diagnostic marker for gynecologic cancers.

6. Generating proteins and peptides encoded by a nucleic acid was routine as of 1998, and there were a number of computer programs routinely used as of 1998 to identify potential open reading frames and deduced proteins and peptides expressed by a defined nucleic acid sequence.

7. Also routine as of 1998 was to utilize the generated proteins and peptides encoded by the defined nucleic acid sequence (such as SEQ ID NO:1 or its fragments) to make antibodies and to routinely select antibodies for their ability to detect cancer.

As discussed below, once the nucleic acid sequence is specified there were several approaches available to those skilled in the art in 1998 to generate antibodies that could be used to formulate tests for circulating proteins originating from the nucleic acid sequence revealed. It is the teaching of the patent that this sequence, and the other sequences revealed by the "ovary specific gene" approaches described, is associated with ovarian cancer that focuses the well established routine work needed to then generate and validate antibody-based diagnostic methods that utilize the coded proteins as biomarkers for ovarian cancer.

8. Further, I disagree with the Examiner's suggestion that identification of a protein sequence or an ORF in the patent application is required for one of skill to identify structural or functional attributes of antibodies to proteins or peptide fragments of a defined nucleic acid sequence.

The nucleic acid sequences contain all the information needed for one skilled in the art to predict, using software tools available in 1998, all proteins that could be coded. These protein sequences could then be used in homology searches, again using software and databases available at the time, to identify target immunogens for specific antibody generation.

The predicted sequences could also have been subjected to antigenic epitope modeling to identify small immunogens that could easily be synthesized and use to generate panels of site specific monoclonal antibodies which then could be routinely selected for recognition of endogenous protein products of the nucleotide sequences taught in the patent.

9. Thus, I believe this patent application does provide sufficient information to identify antibodies or antibody fragments that bind to and/or detect proteins or protein fragments expressed by SEQ ID NO:1 which are useful as cancer diagnostic agents.

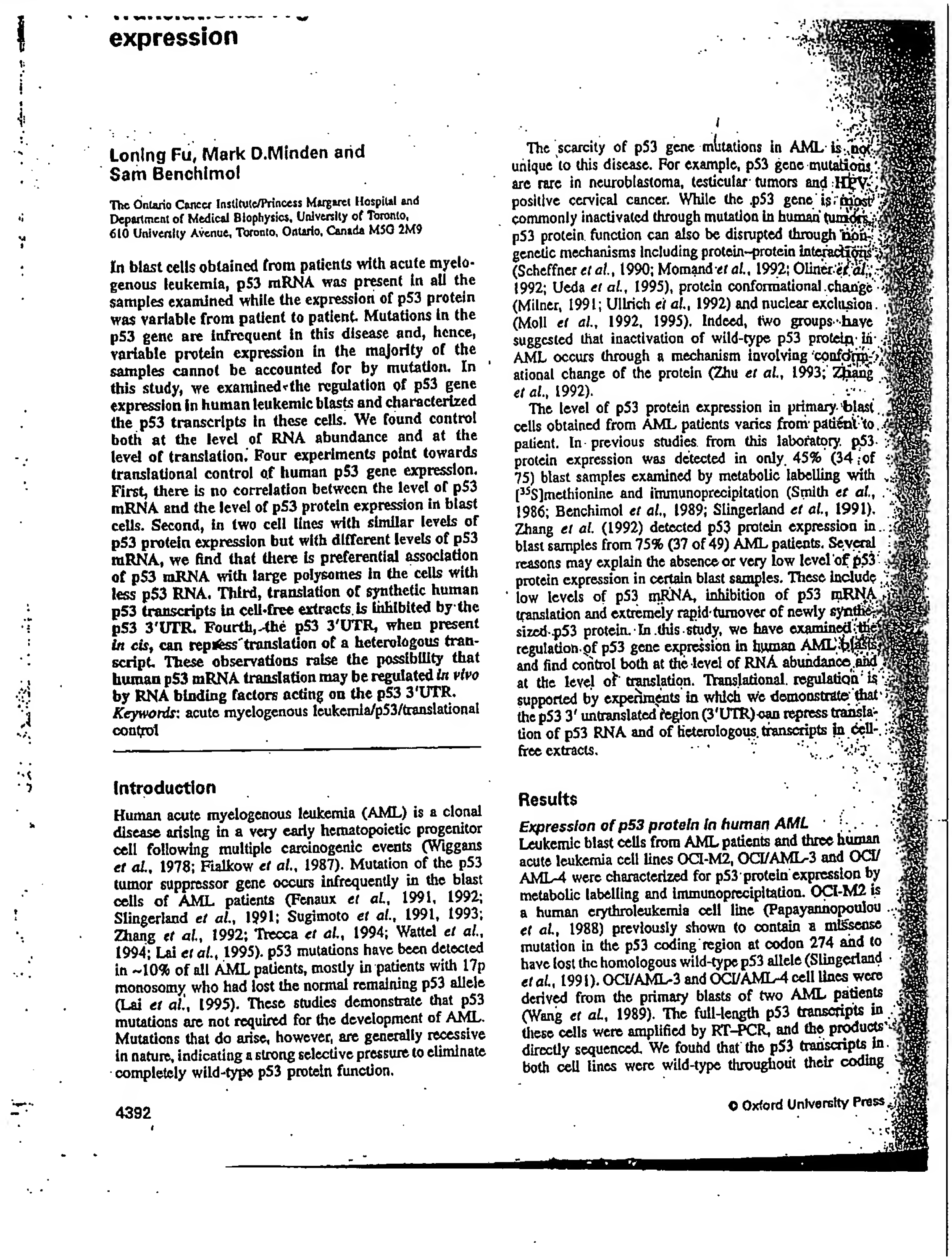
I hereby declare that all statements herein of my own knowledge are true and that all statements made on information or belief are believed to be true; and further that these statements were made with the knowledge that willful statements and the like so made are punishable by fine or by imprisonment, or both, under §1001 of Title 18 of the United States Code, and that such willful statements may jeopardize the validity of the application, any patent issuing there upon, or any patent to which this verified statement is directed.

Patrick M. Sluss

4/21/2008

Date

Appendix D



expression

Loning Fu, Mark D.Minden and Sam Benchimol

The Ontario Cancer Institute/Princess Margaret Hospital and Department of Medical Biophysics, University of Toronto, 610 University Avenue, Toronto, Ontario, Canada M5G 2M9

In blast cells obtained from patients with acute myelogenous leukemia, p53 mRNA was present in all the samples examined while the expression of p53 protein was variable from patient to patient. Mutations in the p53 gene are infrequent in this disease and, hence, variable protein expression in the majority of the samples cannot be accounted for by mutation. In this study, we examined the regulation of p53 gene expression in human leukemic blasts and characterized the p53 transcripts in these cells. We found control both at the level of RNA abundance and at the level of translation. Four experiments point towards translational control of human p53 gene expression. First, there is no correlation between the level of p53 mRNA and the level of p53 protein expression in blast cells. Second, in two cell lines with similar levels of p53 protein expression but with different levels of p53 mRNA, we find that there is preferential association of p53 mRNA with large polysomes in the cells with less p53 RNA. Third, translation of synthetic human p53 transcripts in cell-free extracts is inhibited by the p53 3'UTR. Fourth, the p53 3'UTR, when present *in cis*, can repress translation of a heterologous transcript. These observations raise the possibility that human p53 mRNA translation may be regulated *in vivo* by RNA binding factors acting on the p53 3'UTR.

Keywords: acute myelogenous leukemia/p53/translational control

Introduction

Human acute myelogenous leukemia (AML) is a clonal disease arising in a very early hematopoietic progenitor cell following multiple carcinogenic events (Wiggans *et al.*, 1978; Fialkow *et al.*, 1987). Mutation of the p53 tumor suppressor gene occurs infrequently in the blast cells of AML patients (Fenaux *et al.*, 1991, 1992; Slingerland *et al.*, 1991; Sugimoto *et al.*, 1991, 1993; Zhang *et al.*, 1992; Trecca *et al.*, 1994; Wattel *et al.*, 1994; Lai *et al.*, 1995). p53 mutations have been detected in ~10% of all AML patients, mostly in patients with 17p monosomy who had lost the normal remaining p53 allele (Lai *et al.*, 1995). These studies demonstrate that p53 mutations are not required for the development of AML. Mutations that do arise, however, are generally recessive in nature, indicating a strong selective pressure to eliminate completely wild-type p53 protein function.

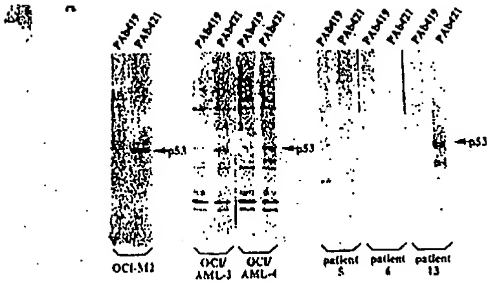
The scarcity of p53 gene mutations in AML is not unique to this disease. For example, p53 gene mutations are rare in neuroblastoma, testicular tumors and HPV-positive cervical cancer. While the p53 gene is most commonly inactivated through mutation in human tumors, p53 protein function can also be disrupted through non-genetic mechanisms including protein–protein interactions (Scheffner *et al.*, 1990; Momand *et al.*, 1992; Oliner *et al.*, 1992; Ueda *et al.*, 1995), protein conformational change (Milner, 1991; Ullrich *et al.*, 1992) and nuclear exclusion (Moll *et al.*, 1992, 1995). Indeed, two groups have suggested that inactivation of wild-type p53 protein in AML occurs through a mechanism involving conformational change of the protein (Zhu *et al.*, 1993; Zhang *et al.*, 1992).

The level of p53 protein expression in primary blast cells obtained from AML patients varies from patient to patient. In previous studies from this laboratory p53 protein expression was detected in only 45% (34 of 75) blast samples examined by metabolic labelling with [^{35}S]methionine and immunoprecipitation (Smith *et al.*, 1986; Benchimol *et al.*, 1989; Slingerland *et al.*, 1991). Zhang *et al.* (1992) detected p53 protein expression in blast samples from 75% (37 of 49) AML patients. Several reasons may explain the absence or very low level of p53 protein expression in certain blast samples. These include low levels of p53 mRNA, inhibition of p53 mRNA translation and extremely rapid turnover of newly synthesized p53 protein. In this study, we have examined the regulation of p53 gene expression in human AML blasts and find control both at the level of RNA abundance and at the level of translation. Translational regulation is supported by experiments in which we demonstrate that the p53 3′ untranslated region (3′UTR) can repress translation of p53 RNA and of heterologous transcripts in cell-free extracts.

Results

Expression of p53 protein in human AML

Leukemic blast cells from AML patients and three human acute leukemia cell lines OCI-M2, OCI/AML-3 and OCI/AML-4 were characterized for p53 protein expression by metabolic labelling and immunoprecipitation. OCI-M2 is a human erythroleukemia cell line (Papayannopoulou *et al.*, 1988) previously shown to contain a missense mutation in the p53 coding region at codon 274 and to have lost the homologous wild-type p53 allele (Slingerland *et al.*, 1991). OCI/AML-3 and OCI/AML-4 cell lines were derived from the primary blasts of two AML patients (Wang *et al.*, 1989). The full-length p53 transcripts in these cells were amplified by RT–PCR, and the products directly sequenced. We found that the p53 transcripts in both cell lines were wild-type throughout their coding



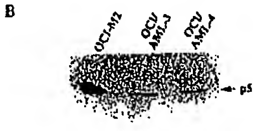


Fig. 1. Expression of p53 protein in human leukemia cells. (A) Cell lines OCI-M2, OCI/AML-3 and OCI/AML-4, and blast cells from AML patients were metabolically labelled with [^{35}S]methionine for 15 min at 37°C. Cell extracts were prepared and portions representing equal amounts of trichloroacetic acid-insoluble radioactivity (10^7 c.p.m.) were immunoprecipitated with the control monoclonal antibody (PAb419) or with monoclonal antibodies against p53 (PAb421). (B) Detection of p53 protein in 5×10^6 cells by Western immunoblotting and ECL using PAb1801 monoclonal antibodies.

regions as well as through their 5′- and 3′UTRs. The only difference detected in the p53 transcripts expressed in OCI/AML-3 and OCI/AML-4 cell lines was the recognized polymorphism at codon 72 (Matlashewski *et al.*, 1987) resulting in an arginine residue in OCI/AML-3 and a proline residue in OCI/AML-4 at position 72.

The level of protein expression measured by metabolic labelling and immunoprecipitation is dependent primarily on the rate of protein synthesis, the rate of protein degradation and the amount of mRNA available for translation. To minimize the contribution of protein half-life on the detection of p53 protein synthesis during the metabolic labelling assay, cells were exposed to a short 15 min pulse of [^{35}S]methionine at 37°C followed by immediate lysis on ice in the presence of protease inhibitors. Radiolabelled cell extracts prepared in this way were then subjected to immunoprecipitation with p53-specific antibodies. p53 protein with a half-life much less than 15 min, however, might remain undetectable by this assay. p53 protein synthesis was detected in OCI/AML-3, OCI/AML-4 and in OCI-M2 (Figure 1A) as well as in seven of 16 blast samples tested; three representative examples are shown in Figure 1A.

The steady-state level of p53 protein in the three cell lines was determined by Western blot analysis using PAb1801. Densitometric scanning of the blot shown in Figure 1B revealed that the amount of p53 protein in OCI/AML-3 and OCI/AML-4 was similar and ~10-fold lower than in OCI-M2. The high level of p53 protein in OCI-M2 was expected since mutant p53 polypeptides usually have much longer half-lives than wild-type p53 proteins

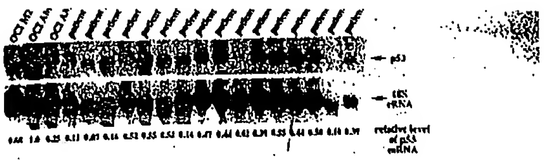


Fig. 2. Northern blot analysis of p53 mRNA in human AML cells. 20 μg of total RNA isolated from cell lines or patient blast samples was separated on a 1% agarose gel containing 6% formaldehyde, transferred to nitrocellulose and hybridized with ^{32}P-labelled human p53 cDNA. After autoradiography, the probe was removed and the filters were hybridized with a probe specific for 18S ribosomal RNA. The relative abundance of p53 mRNA was determined by phosphorimage analysis after normalizing to the value of 18S ribosomal RNA in each sample.

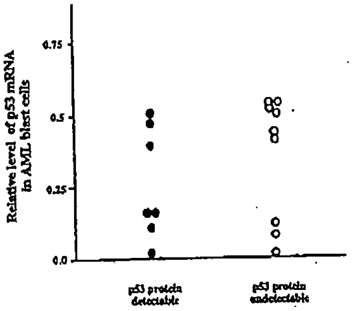


Fig. 3. Relative abundance of p53 mRNA in cells that do or do not express detectable p53 protein. p53 protein synthesis was assessed in 16 AML blast samples by metabolic labelling with [^{35}S]methionine for 15 min and immunoprecipitation. p53 protein synthesis was detected in seven of these samples. p53 mRNA levels were determined by Northern blot analysis as described in the legend to Figure 2.

and as a result mutant p53 polypeptides accumulate intracellularly.

Expression of p53 mRNA in human AML

To determine whether the differences in p53 protein expression in leukemic blasts reflected differences in the abundance of p53 mRNA, RNA was isolated from AML blast samples and cell lines, and subjected to Northern blot analysis. The relative abundance of p53 mRNA in cells was estimated by phosphorimage analysis after normalizing to the value of 18S ribosomal RNA in each sample. The results are shown in Figure 2 and indicate that the 16 AML blast samples examined synthesized a single species of full-length p53 mRNA ~2.8 kb in size. The relative amount of p53 mRNA in the 16 samples varied over a 27-fold range. No correlation was evident between p53 protein expression (on the basis of the 15-min metabolic labelling assay) and the level of p53 mRNA in AML blasts (Figure 3).

OCI/AML-3 and OCI/AML-4 cells contained similar amounts of p53 protein. However, the RNA blot shown in Figure 2 indicated that the abundance of p53 mRNA was 4-fold higher in OCI/AML-3 than in OCI/AML-4. A 4- to 8-fold difference in p53 RNA was seen in repeated

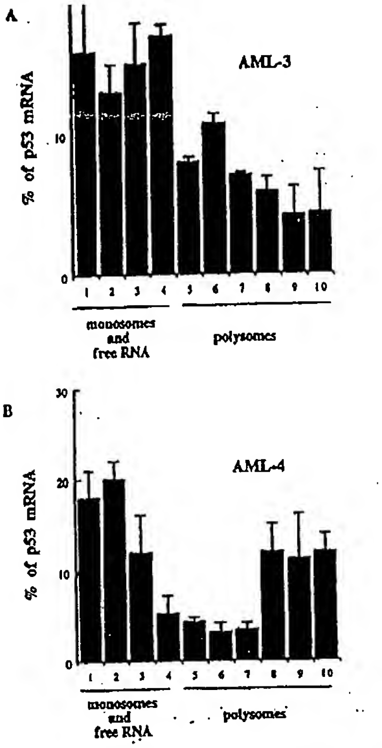


Fig. 4. Association of p53 mRNA with polysomes in OCI/AML-3 (A) and OCI/AML-4 (B) cells. The association of p53 mRNA with polysomes in OCI/AML-3 and OCI/AML-4 cells was compared. Cell extracts containing polysomes were prepared in the presence of cycloheximide and loaded on a 15–50% linear sucrose gradient. Ten fractions were collected and the amount of p53 mRNA in each fraction was determined by dot-blot hybridization analysis with a ^{32}P-labelled human p53 cDNA probe. The size of the polysomes with respect to the gradient was estimated using a polysome preparation from OCI/AML-3 cells. The positions of free ribosomes, monosomes and polysomes are indicated. Error bars represent the standard error of the mean from three separate experiments.

experiments after normalization with probes that detect 18S ribosomal RNA or GAPDH to ensure equivalent loading of RNA samples on the gels. We conclude that p53 RNA levels and p53 protein expression are variable in AML blasts and cell lines, and that the level of p53 protein expression is not related to the amount of p53 mRNA in these cells.

Association of p53 mRNA with polysomes

To test whether p53 gene expression is under translational control *in vivo*, the association of p53 mRNA with polysomes in OCI/AML-3 and OCI/AML-4 cells was analyzed (Figure 4). If p53 mRNA is more translationally active in OCI/AML-4 than in OCI/AML-3 as the above results suggest, then a larger proportion of the p53 mRNA polysomes compared with OCI/AML-3. Cells were collected and lysed in the presence of cycloheximide and MgCl$_2$, which stabilize the association of ribosomes with mRNA. The lysates were sedimented through a linear sucrose gradient and fractions were collected. RNA was extracted from each fraction and analyzed for the presence of p53 mRNA by dot-blot hybridization with a ^{32}P-labelled p53 cDNA probe. The gradients were calibrated with polysomes prepared from lysates by precipitation with 100 mM MgCl$_2$. Polysomes were found at the bottom of the gradient in fractions 5–10, while monosomes were found in fractions 1–4. p53 mRNA from OCI/AML-4 cells was associated with larger polysomes than was p53 mRNA from OCI/AML-3 cells (Figure 4). In OCI/AML-4 cells, 39% of the p53 mRNA was found in fractions 7–10 containing high molecular weight polysomes, while in OCI/AML-3 cells 21% of the p53 mRNA was found in these same fractions. As an internal control, the distribution of ribosomal protein L35 RNA was compared and shown to be identical in OCI/AML-3 and OCI/AML-4 (data not shown).

Analysis of the 5′ end of p53 mRNA

The human p53 gene has been shown to have a cluster of six or seven major transcription initiation sites and several minor sites lying further upstream (Tuck and Crawford, 1989). Transcripts initiating from the minor sites would have a longer 5′UTR with potential to form a stable stem–loop structure close to the 5′ cap. Such structures would not be expected to form in transcripts initiating from the major start sites. 5′-stem–loop structures were described for rodent p53 mRNA (Bienz *et al.*, 1984; Bienz-Tadmor *et al.*, 1985). Recently, mouse p53 protein was shown to bind to the 5′UTR and to inhibit translation of its own mRNA in an *in vitro* assay system (Mosner *et al.*, 1995). Stable stem–loop structures in the 5′UTR regions of a number of mRNA transcripts have been shown to inhibit translation initiation by interfering with the activity of translation initiation factors or by serving as binding sites for regulatory proteins that inhibit translation (Feng and Holland, 1988; Fu *et al.*, 1991; Melefors and Hentze, 1993; Pause *et al.*, 1993).

To determine if the low level of p53 protein expression in leukemic blasts was the result of transcription initiating at the minor start sites, the 5′ ends of p53 mRNA present in different blast samples and cell lines were mapped using an RNase protection assay. A 729 nucleotide antisense RNA probe containing genomic sequences from the p53 promoter region fused with cDNA sequences extending into exon 4 was generated by transcription with SP6 RNA polymerase in the presence of [^{32}P]UTP (Figure 5A). This probe would yield protected p53 fragments of 385 nucleotides corresponding to transcripts originating from the major start site and 449 nucleotides corresponding to transcripts originating from the most 5′ of the minor start sites. Total RNA extracted from OCI/AML-3 and OCI/AML-4 cell lines and from seven AML blast samples was examined. After digestion, the protected fragments were resolved by electrophoresis on a denaturing polyacrylamide gel. As shown in Figure 5B, the predominant protected fragment in all the RNA samples was 385 nucleotides in length indicating a common site for initiation

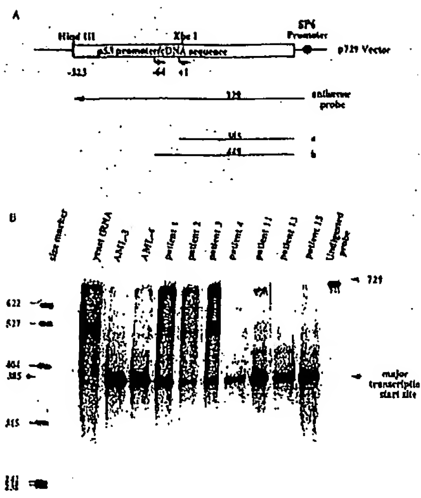


Fig. 5. RNase protection assay. (A) The map of the p729 plasmid. The p729 plasmid was constructed as described under Materials and methods. After linearization with *Hin*dIII, a 729 nucleotide antisense RNA probe was generated by transcription with SP6 RNA polymerase yielding protected p53 fragments of ~385 nucleotides due to p53 transcripts initiating from one of the major start sites (a) and 449 nucleotides due to p53 transcripts initiating from the most 5′ of.the minor transcription start sites (b). (B) The 729 nucleotide [^{32}P]UTP-labelled antisense RNA probe was annealed to 30 μg of total RNA extracted from OCI/AML-3 and OCI/AML-4 cell lines and seven AML blast samples before digestion with RNase A and RNase T1. The protected fragments were separated by electrophoresis on a 6% polyacrylamide–8 M urea gel and visualized by autoradiography. The positions and size (nucleotide length) of 5′ end-labelled fragments of *Msp*I-digested pBR322 plasmid DNA are indicated on the left. The bottom arrow indicates the position of the major protected fragment and the top arrow indicates the undigested probe.

of p53 gene transcription in leukemic blasts at the major start site. These data indicate that, in contrast with murine p53 mRNA, stable secondary structures are unlikely to exist at the 5′ end of human p53 mRNA.

Analysis of the 3′ end of p53 mRNA

Human p53 mRNA contains a long 3′UTR of 1176 nucleotides with an Alu-like repetitive sequence element of ~470 bp located immediately upstream of the poly(A) tail (Matlashewski *et al.*, 1984). The Alu-like sequence is in the reverse transcriptional orientation with respect to the p53 gene. Furthermore, the Alu-like sequence is missing in murine p53 transcripts and it interrupts a region in human p53 mRNA which shows homology to mouse p53 mRNA. When analyzed with the FOLD program of GCG, the Alu-like element in the 3′UTR of human p53 mRNA is predicted to form an independent secondary structure that does not have long-range interactions with other regions of p53 mRNA. In the presence of a poly(A) tail, the secondary structure formed by the Alu-like element is predicted to remain essentially intact except that a 50 nucleotide U-rich sequence at the 5′ boundary of the Alu-like sequence will interact with the poly(A) tail. The extended base pairing between U and A residues will further stabilize the secondary structure formed by the Alu-like element. To determine whether or not the Alu-like repeat present in human p53 mRNA might constitute a negative regulatory element during translation, a series of *in vitro* transcription–translation experiments was performed.

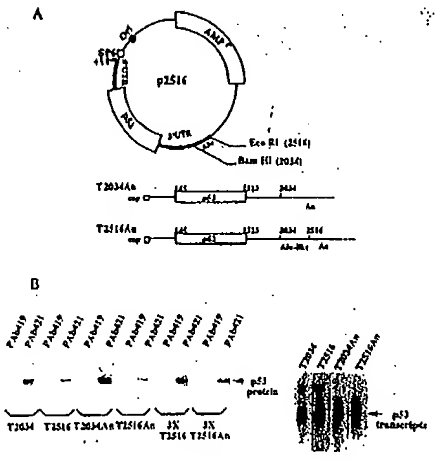


Fig. 6. *In vitro* translation of synthetic p53 RNA containing variable portions of the 3′UTR. (A) Plasmid template used to synthesize p53 RNA *in vitro*. The p2516 plasmid was constructed by inserting the entire 2.5 kb wild-type p53 cDNA sequence downstream of the bacteriophage SP6 promoter in a pSP64-derived plasmid. Transcription from the SP6 promoter present in p2516 leads to the production of transcripts in which the first 10 nucleotides are derived from plasmid sequences while the remaining nucleotides are derived from the p53 gene beginning at the +7 position of native p53 transcripts initiating from one of the major transcription start sites (Tuck and Crawford, 1989). Linearization of p2516 at the *Eco*RI site before *in vitro* transcription generates a full-length, Alu-containing p53 transcript (T2516); linearization at the *Bam*HI site provides a template for the synthesis of a truncated p53 transcript missing a portion of the 3′UTR containing the Alu sequence (T2034). Both transcripts were polyadenylated *in vitro* to generate p2516An and p2034An. The open rectangles shown on the transcripts represent the position of the p53 coding region. (B) 50 ng of the *in vitro*-synthesized T2034, T2516, T2034An and T2516An p53 RNAs were translated in a rabbit reticulocyte lysate at 30°C for 30 min in the presence of [^{35}S]methionine followed by immunoprecipitation, SDS–PAGE and autoradiography. In the 3X T2516 and 3X T2516An lanes, 150 ng of T2516 or T2516An RNA was added to the *in vitro* translation reaction. The right panel presents the results of a Northern blot in which 50 ng of synthetic p53 RNA was applied to an agarose–formaldehyde gel, blotted and hybridized to ^{32}P-labelled human p53 cDNA.

An SP6-derived plasmid containing human wild-type p53 cDNA including the entire 3′UTR was constructed (p2516 in Figure 6A). p2516 was linearized with *Eco*RI or with *Bam*HI and used as a template for *in vitro* transcription. In some reactions, a poly(A) tail of 200–300 adenylic acid residues was added to synthetic p53 RNA using poly(A) polymerase. In this way, four synthetic p53 transcripts were generated: T2516An and T2516 represent full-length, Alu-containing transcripts with or without a poly(A) tail; T2034An and T2034 represent poly(A) tail. These transcripts were then used as templates for translation in a rabbit reticulocyte lysate containing [^{35}S]methionine. p53 protein synthesized *in vitro* was immunoprecipitated with PAb421 monoclonal antibody and visualized by autoradiography (Figure 6B). The amount and integrity of the synthetic p53 RNAs added to the *in vitro* translation reactions was monitored by agarose gel electrophoresis and Northern blotting as shown in the right panel of Figure 6B. Densitometric tracing of the data indicated that the Alu-containing, non-polyadenylated transcript T2516 was translated ~3-fold less efficiently than the Alu-deficient, non-polyadenylated transcript T2034. In addition, the polyadenylated, Alu-containing transcript T2516An was translated ~20-fold less efficiently than the polyadenylated, Alu-deficient transcript T2034An. These data indicate that the Alu-like element present in the p53 3'UTR can inhibit p53 mRNA translation *in vitro*, even in the absence of a poly(A) tail. The predicted interaction of the poly(A) tail with the Alu-like element appears to increase further the inhibition of translation.

To test further the inhibitory activity of the p53 3'UTR, we examined the ability of the p53 3'UTR to control the translation of a heterologous RNA. The Alu-containing p53 DNA fragment extending from nucleotides 2034 to 2516 was excised from plasmid p2516 and inserted downstream of a heterologous gene (CAT gene) in an SP6-based plasmid vector to generate the plasmid pCAT-Alu (Figure 7A). *In vitro* transcription and translation revealed that non-polyadenylated CAT-Alu RNA was translated 5-fold less efficiently than non-polyadenylated CAT transcripts lacking the Alu sequence (Figure 7B). When a different region of the p53 3'UTR (nucleotides 1465–2034 in plasmid p2516) with approximately the same length as the Alu-containing fragment was inserted downstream of the CAT gene, no effect on CAT translation was observed (CAT-BS in Figure 7B). The ability of the Alu-containing segment of the p53 3'UTR to act on a heterologous transcript indicates that it likely represses translation independently of upstream sequences.

The inhibitory activity of the Alu-like element on p53 translation was likely the result of its action *in cis* and not simply due to non-specific inhibition of translation, since a 3-fold increase in the amount of Alu-containing transcript added to the reticulocyte lysate resulted in a corresponding increase in the amount of p53 protein synthesized (Figure 6B). Furthermore, when 200 ng of luciferase RNA was added to a reticulocyte lysate together with 200 ng of CAT-Alu or CAT-BS RNA, there was little difference in the amount of luciferase synthesized (Figure 7C). Similarly, when 200 ng of luciferase RNA was added to a reticulocyte lysate, either alone or mixed with 200 ng of T2034 or T2516An RNA, there was little difference in the amount of luciferase synthesized (data not shown).

To confirm that the decrease in p53 protein synthesis from Alu-containing p53 RNAs was due to translational regulation and not due to preferential RNA degradation in the reticulocyte lysate, adenylated T2034 and T2516 synthetic transcripts were added to the rabbit reticulocyte lysate under the same conditions as those used for *in vitro* translation. After incubation for 15 or 60 min, RNA was extracted from the lysate and the amount of synthetic p53 RNA present in the lysate determined by Northern blot analysis. Enhanced degradation of the Alu-containing transcript was not observed (Figure 8). We conclude that a segment of the p53 3'UTR encompassing the Alu-like element is capable of repressing translation *in vitro*.

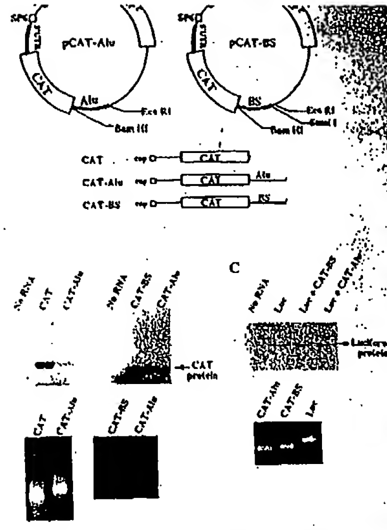


Fig. 7. The p53 Alu-like element can inhibit translation of a heterologous CAT transcript. (A) Plasmids used to generate CAT transcripts *in vitro*. (B) 200 ng of *in vitro*-synthesized CAT, CAT-Alu, and CAT-BS transcripts were translated in a rabbit reticulocyte lysate at 30°C for 30 min in the presence of [^{35}S]methionine. The reactions were stopped by adding an equal volume of the 2X protein sample buffer, heated to 100°C for 5 min and analyzed by SDS–PAGE and autoradiography. An ethidium bromide-stained agarose gel demonstrating the integrity and amount of synthetic transcripts that were added to the *in vitro* translation reaction is shown below. (C) 200 ng of luciferase RNA was translated in a rabbit reticulocyte lysate either alone or in the presence of 200 ng of CAT-BS or 200 ng of CAT-Alu. Reaction mixtures were incubated in the presence of [^{35}S]methionine at 30°C for 30 min and processed as in (B). The *in vitro*-synthesized luciferase protein is shown in the upper panel; the RNA used for *in vitro* translation is shown in the ethidium bromide stained-agarose gel in the bottom panel.

Discussion

The observation that wild-type p53 protein expression in leukemic blast cells does not correlate with the level of p53 mRNA mirrors findings reported previously for blasts and other human cell types (Matlashewski *et al.*, 1986; Kastan *et al.*, 1991a; Slingerland *et al.*, 1991; Sasano *et al.*, 1992; Hsu *et al.*, 1993). The absence of detectable p53 protein in cells expressing abundant levels of wild-type p53 mRNA has usually been attributed to the short half-life of p53 protein in normal cells (Rogel *et al.*, 1985). A similar situation exists in papillomavirus (HPV)-infected cells such as HeLa cells where p53 protein is not detected even though these cells produce p53 mRNA and this RNA is associated with polysomes (Matlashewski

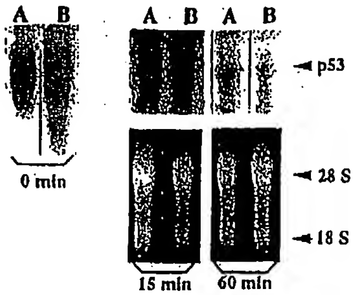


Fig. 8. Stability of synthetic human p53 RNAs in rabbit reticulocyte lysates. 100 ng of adenylated T2034 (A) and T2516 (B) synthetic RNA was added to the rabbit reticulocyte lysate and incubated at 30°C for 15 or 60 min under the same conditions used for *in vitro* translation. RNA present in the lysates was then extracted and loaded on a 1% agarose-formaldehyde gel. The 0 min time point represents 100 ng of synthetic RNA loaded directly on the gel. The amount of p53 RNA in each sample was then determined by Northern blotting using a ^{32}P-labelled human p53 cDNA. The lower panel shows the 28S and 18S ribosomal RNAs recovered from the rabbit reticulocyte lysates detected by ethidium bromide staining of the gel.

et al., 1986). The enhanced degradation of newly synthesized p53 protein in HeLa cells was shown to be promoted by the papillomavirus E6 protein which is expressed constitutively in these cells (Scheffner *et al.*, 1990).

In this report, we present data showing that differences in p53 mRNA abundance exist in AML blasts and that these differences cannot explain the heterogeneity in the level of p53 protein expression in leukemic blast cells. Using a metabolic labelling assay in which blasts from different AML patients were pulse-labelled with [^{35}S]-methionine for 15 min to minimize the contribution of protein half-life on the detection of p53 protein synthesis, we found differences in the level of p53 protein expression in blast samples. These observations raised the possibility that p53 gene expression may be regulated at the translational level in certain human cells. We tested this possibility by analyzing the distribution of p53 mRNA on polysomes *in vivo* and by examining p53 RNA translation *in vitro*.

We have used two AML cell lines, OCI/AML-3 and OCI/AML-4 that contain similar amounts of wild-type p53 protein even though OCI/AML-3 contains 4- to 8-fold more p53 mRNA. Comparison of the polysome profile of these cells indicated that a greater proportion of the p53 mRNA was associated with larger polysomes in OCI/AML-4 than in OCI/AML-3. p53 mRNA in both of these cell lines as well as in blasts from different AML patients is present as a single, full-length species of ~2.8 kb that initiates from a common transcription start site and contains similar sequence and structural elements.

Transcription–translation experiments *in vitro* indicated that the p53 3'UTR contains a negative regulatory domain that is capable of repressing translation *in vitro*. A region of the 3'UTR consisting of ~500 nucleotides and containing an Alu-like element is capable of repressing translation of p53 mRNA and of a heterologous transcript. The p53 3'UTR, when present *in cis*, repressed translation of polyadenylated as well as non-polyadenylated transcripts. Accordingly, we suggest that the Alu-like element, possibly through its secondary structure, is capable of repressing p53 mRNA translation. In addition, interaction of the Alu-like element with the poly(A) tail may repress the latter's function in translation. Experiments are in progress to map precisely this regulatory element in the p53 3'UTR and to determine if the p53 3'UTR plays a similar role in regulating translation *in vivo*.

Our finding that p53 protein expression in AML blasts is controlled, at least in part, through mechanisms acting at the translational level, raises the possibility that translational regulation may provide an epigenetic mechanism to reduce or even eliminate wild-type p53 protein function in leukemic blasts. In preliminary experiments to address this point, we have exposed blast cells that express little or no detectable p53 protein to 6 Gy of ionizing radiation and have observed increased steady-state levels of p53 protein at 1.5 h after irradiation (data not shown). Genotoxic agents have been shown previously to increase the level and/or activity of p53 protein through a post-transcriptional mechanism that is not well understood (Kastan *et al.*, 1991b; Fritsche *et al.*, 1993; Lu and Lane, 1993; Zhan *et al.*, 1993). Hence, blast cells retain the ability to up-regulate p53 expression in response to genotoxic stress. At least under these conditions, p53 function may not be lost. This type of analysis, however, does not address the function of p53 in proliferating cells that have not been exposed to genotoxic stress. In this regard, previous studies from our laboratory demonstrated a highly significant correlation between p53 protein expression in leukemic blast cells and the secondary plating efficiency of these cells (Smith *et al.*, 1986). The latter provides an estimate of the self-renewal capacity of progenitor cells in the blast population. Deregulated p53 expression might, therefore, be expected to affect the self-renewal capacity of blasts in the absence of genotoxic stress.

Accumulating evidence demonstrates the involvement of the 3'UTR in translational control (Jackson, 1993). The demonstration that the 3'UTR of certain transcripts can control mRNA localization and polyadenylation provides a mechanism for translational regulation (Huarte *et al.*, 1992; Gavis and Lehmann, 1994). In addition, specific sequences within 3'UTRs have been shown to repress translation (Goodwin *et al.*, 1993; Evans *et al.*, 1994; Kwon and Hecht, 1993). RNA–protein interactions are likely to be involved in 3'UTR-dependent translational repression. Indeed, a protein that binds specifically to the 3'UTR of protamine 2 mRNA and represses its translation has been identified (Kwon and Hecht, 1993). If the p53 3'UTR can be shown to regulate p53 mRNA translation *in vivo*, it is possible that *trans*-acting factors (missing or inactive in reticulocyte lysates) activate components of the translational machinery to bypass this negative regulatory domain on human p53 mRNA. Such *trans*-acting factors could interact directly with p53 mRNA to enhance its rate of translation. Alternatively, *trans*-acting factors directed to the p53 3'UTR (that are also present in reticulocyte lysates) may act as repressors of translation. Differences in the level of p53 protein synthesis among AML blasts and possibly other human cells could, therefore, be determined by differences in the level or activity of these regulatory molecules.

Materials and methods

Cells

The OCI/AML-3 and OCI/AML-4 cell lines were derived from primary blasts of two AML patients (Wang *et al.*, 1989). The OCI-M2 cell line was derived from the primary blasts of a patient whose erythroleukemia represented the end stage of a previously identified myelodysplastic syndrome (Papayannopoulou *et al.*, 1988). OCI/AML-3 and OCI-M2 cells were grown in alpha-modified minimum essential medium (α-MEM) containing 10% fetal calf serum (FCS) (GIBCO). The OCI/AML-4 cells were grown in α-MEM containing 10% FCS and 10% conditioned medium obtained from the human bladder carcinoma cell line 5637 (5637-CM) (Wang *et al.*, 1989). The AML blast cells were obtained directly from AML patients. The mononuclear cell fraction of peripheral blood was collected after separation through Ficoll–Hypaque (Pharmacia) (1.077 g/ml) and T-lymphocyte depletion (Minden *et al.*, 1979). These cells were stored frozen in liquid nitrogen before use.

Metabolic labelling and immunoprecipitation

The blast cells of AML patients were thawed and incubated for 2 days at 37°C in α-MEM containing 10% FCS and 10% 5637-CM before metabolic labelling. 1×10^7 cells were labelled with 0.2 mCi $[^{35}S]$-methionine (DuPont NEN Research Products) in 0.5 ml α-MEM lacking methionine and containing 10% dialysed FCS at 37°C for 15 min. Cells were then immediately pelleted, the radioactive medium removed, and the cells lysed on ice in a solution containing 25 mM Tris pH 7.4, 50 mM NaCl, 0.5% sodium deoxycholate, 2% NP40, 0.2% SDS, 0.5 mM phenylmethylsulfonyl fluoride (PMSF), 1 μg/ml leupeptin and 1 μg/ml aprotinin for 20 min. Lysates were cleared by centrifugation, the supernatant was retained and incubated with 5 μg of a non-specific IgG2a mouse monoclonal antibody (Sigma) for 60 min on ice. These were then reacted with 0.5 ml of a 10% suspension of formalin-treated *Staphylococcus aureus* Cowan I cells (Pansorbin, Calbiochem-Behring) for 30 min on ice, followed by centrifugation and retention of the supernatant. Portions of precleared lysates containing equal numbers of trichloroacetic acid-insoluble counts (10^7 c.p.m.) were diluted in NET/GEL buffer (150 mM NaCl, 5 mM EDTA pH 8.0, 50 mM Tris pH 7.4, 0.05% NP40, 0.02% sodium azide, 0.25% gelatin) and immunoprecipitated on ice for 2 h with PAb421 monoclonal antibodies against p53 protein or control PAb419 antibodies (Harlow *et al.*, 1981). The immune complexes were collected on 60 μl prewashed protein A–Sepharose beads (Pharmacia), washed three times with NET/GEL buffer, and eluted into 30 μl protein sample buffer (2% SDS, 10% glycerol, 0.1% bromophenol blue, 25 mM Tris pH 6.8, 0.1 M dithiothreitol) by boiling for 10 min. The Sepharose beads were removed by centrifugation, the samples were loaded on a 10% polyacrylamide gel containing SDS and proteins were resolved by electrophoresis at 45 mA. Gels were fixed in 7.5% acetic acid and 25% methanol for 30 min before drying and exposure to X-ray film (DuPont NEN Research Products).

Western blot analysis

5×10^6 cells were lysed directly in an equal volume of 2X protein sample buffer. The extracts were passed through a 21-gauge needle several times to reduce viscosity and boiled for 10 min before electrophoresis at 45 mA on a 10% polyacrylamide gel containing SDS. Resolved proteins were transferred to a nitrocellulose membrane (Schleicher & Schuell), and the abundance of p53 protein was estimated by immunoblotting with a human p53-specific monoclonal antibody PAb1801 (Banks *et al.*, 1986). Bound antibody was detected using the enhanced chemiluminescence detection system (DuPont NEN Research Products) according to the manufacturer's instructions.

Northern blot analysis

Total cellular RNA was isolated using the guanidinium thiocyanate–cesium chloride method (Chirgwin *et al.*, 1979). 20 μg of total RNA was separated by electrophoresis on a 1% agarose gel containing 6% formaldehyde and transferred to a nitrocellulose membrane (Schleicher & Schuell). The blots were hybridized with cDNA probes labelled with $[^{32}P]$dCTP in a random priming reaction (Feinberg and Vogelstein, 1983), washed and exposed to X-ray film. The amount of RNA was determined with a Molecular Dynamics PhosphorImager using Multiquant software. The human p53 probe was the *Xba*I–*Eco*RI fragment of p53 cDNA from the pR4-2 plasmid (Harlow *et al.*, 1985); the L35 probe was the *Pst*I–*Bam*HI fragment from the human ribosomal protein L35 cDNA (Herzog *et al.*, 1990); the GAPDH probe was a 1.3 kb *Pst*I fragment of rat GAPDH cDNA (Fort *et al.*, 1985); the 18S ribosomal gene (Torczynski *et al.*, 1985).

Genomic DNA preparation

Genomic DNA from OCI/AML-3 and OCI/AML-4 cell lines was isolated following a modification of the procedure described by Kuplec *et al.* (1987). 3×10^7 cells were washed with ice-cold PBS buffer, resuspended in 3 ml of lysis buffer (20 mM EDTA pH 8.0, 100 μg/ml proteinase K, 0.5% sarkosyl) and incubated at 50°C for 3 h. DNA was extracted with phenol/chloroform, dialysed against 50 mM Tris–HCl pH 8.0, 10 mM EDTA, 10 mM NaCl at 4°C, and then treated with RNase A (100 μg/ml) at 37°C for 3 h. DNA was again extracted with phenol/chloroform and dialysed against 10 mM Tris pH 7.4, 1 mM EDTA. DNA concentration was determined by measuring the absorbance at 260 nm.

Amplification of p53 sequences from RNA and DNA

20 μg of total RNA was precipitated with ethanol and resuspended in a 30 μl reaction containing 300 ng of oligo(dT) primer (Amersham International), 50 mM Tris–HCl pH 8.3, 77 mM KCl, 3 mM $MgCl_2$, 3 mM dithiothreitol, 3 mM dNTP, 30 units of RNAguard (Pharmacia) and 200 units of Moloney murine leukemia virus reverse transcriptase (GIBCO-BRL) and incubated at 42°C for 60 min. The first strand cDNA was then used as the template for amplification by PCR using *Taq* polymerase (Promega). PCR amplification was performed with 10 μl of each first stand cDNA as the template and 40 cycles of denaturation (94°C, 1 min), annealing (64°C, 30 s), and elongation (72°C, 1 min). The following p53-specific primers were used for amplifying the complete coding region and the 3'UTR: 5'SX1 (sense, exon 1, GACACTTTGCGTTCGGGCTGGGAG), 5'SX5A (sense, exon 5, GAGCGCTGCTCAGATAGCGATG), 3'SX11 (sense, exon 11, GAAGGGCCTGACTCAGACTGAC), 3'AX-6 (antisense, exon 6, AGATGCTGAGGAGGGGCCAGAC), JS-3 (antisense, exon 11, GAGGGAGAGATGGGGGTGGGAGGCTGTC) and AS-4 (antisense, exon 11, GGCAGCAAAGTTTTATTGTAAAATAAG). The 5'UTR and sequences further upstream were amplified from 1 μg genomic DNA using the following pair of p53-specific primers: 5'UTR-1 (sense, promoter region, ACCTAAGCTTGTCATGGCGACTGTCCAGCTTTG) and p-EX (antisense, exon 1, CCAATCCAGGGAAGCGTGTCACCG).

Direct sequencing of double-stranded PCR products

Double-stranded DNA fragments produced by PCR amplification were eluted from agarose gels and purified by extraction with phenol/chloroform. 200 ng of purified PCR product were mixed with human p53-specific oligonucleotides as sequencing primers, frozen in dry ice, dried in a centrifugal evaporator (Savant SpeedVac), redissolved in sequencing buffer (40 mM Tris–HCl pH 7.5, 25 mM $MgCl_2$, 50 mM NaCl, 10% DMSO) and subjected to the sequencing reaction as described by Winship (1989).

RNase protection assay

Plasmid p729 was constructed from three DNA fragments in two stages. A 330 bp DNA fragment derived from the human p53 gene promoter was excised from the p2E-H2BX plasmid (Lamb and Crawford, 1986) with *Hin*dIII and *Xba*I and inserted into the pGEM-4 plasmid (Promega) between the *Hin*dIII and *Xba*I sites. In the second stage, a fragment corresponding to the 5' end of p53 mRNA was obtained by RT–PCR using p53 mRNA prepared from OCI/AML-3 cells and the p53-specific primers 5'UTR-3 (sense, exon 1, CCGGAAGCTTCAAAAGTCTAGAGCCACCGTCCAG) and 5'AX4 (antisense, exon 4, GGTGTAGGAGCTGCTGCTGGTGC). The resulting fragment was end-filled with the Klenow fragment of DNA polymerase I, digested with *Xba*I at the site present in the 5'UTR-3 primer shown underlined and inserted between the *Xba*I and *Sma*I sites present in the plasmid generated in the first stage.

p729 was linearized with *Hin*dIII and a 729 nucleotide antisense probe was prepared by transcription with SP6 RNA polymerase. The *in vitro* transcription reaction mixture contained 50 mM Tris–HCl pH 8.0, 10 mM $MgCl_2$, 4 mM spermidine, 10 mM NaCl, 0.5 mM each of ATP, GTP, CTP, 12 μM UTP, 5 μCi $[^{32}P]$UTP, 10 mM dithiothreitol, 20 units of RNAguard, 0.5 μg of linearized template and 10 units of SP6 RNA polymerase in a final volume of 20 μl. After incubation at 37°C for 60 min, the DNA template was digested with DNase I and the RNA probe was extracted with phenol/chloroform, precipitated with ethanol and resuspended in water. This RNA probe covered the entire p53 gene promoter region and included the first three exons and a part of the fourth exon. p53 transcripts initiating from one of the major start sites should yield protected fragments of ~385 nucleotides: p53 transcripts originating from the most 5' of the minor start sites should yield protected fragments of 449 nucleotides (Tuck and Crawford, 1989).

In the RNase protection assay, 30 μg of total RNA was mixed with 1×10^5 c.p.m. of the labelled probe and precipitated with ethanol. The RNA/probe mixtures were then washed, dried and resuspended in 10 μl of hybridization solution (Winter *et al.*, 1985), heated to 80°C for 10 min, and hybridized at 46°C overnight. After hybridization, the samples were mixed with 0.18 ml of RNase digestion mix containing 60 μg/ml of RNase A (type III, Sigma), 1100 U/ml of RNase T1 (Boehringer Mannheim) in 300 mM NaCl, 5 mM EDTA, 10 mM Tris–HCl pH 7.5. After incubation at 37°C for 60 min, the digestion was terminated by addition of 10 μl of 20% SDS and 5 μl of proteinase K (10 mg/ml) (Boehringer Mannheim) and incubation at 37°C for 15 min. Protected fragments were extracted with phenol/chloroform, precipitated with ethanol, resolved by denaturing gel electrophoresis and visualized by autoradiography.

Polysome analysis

5×10^7 cells were washed once in ice-cold Tris–saline solution (25 mM Tris–HCl pH 7.5, 25 mM NaCl) containing 10 mM MgCl$_2$ and 10 μg/ml cycloheximide. The cells were then immediately lysed on ice with the use of a Dounce homogenizer in 2 ml homogenization buffer containing 25 mM Tris–HCl pH 7.5, 25 mM NaCl, 10 mM MgCl$_2$, 2% Triton X-100, 340 U/ml heparin (LEO Laboratories Canada Ltd), 2 mM vanadyl ribonucleoside complex (Sigma), 2.5 mM PMSF, 10 μg/ml cycloheximide, 1 mM dithiothreitol and 1 mM EGTA. The extract was centrifuged at 14 000 r.p.m. for 6 min at 4°C to remove cell debris, the supernatant was collected and layered over a 15–50% linear sucrose gradient (11 ml) prepared in homogenization buffer. The gradients were centrifuged in an SW41 Beckman rotor at 175 000 *g* for 110 min at 4°C. Ten fractions of equal volume were collected from the bottom of the tubes. RNA was prepared from each of the fractions by phenol/chloroform extraction and ethanol precipitation and resuspended in 200 μl DEPC-treated water. The amount of p53 mRNA in each fraction (100 μl of the RNA sample) was determined by dot-blot hybridization analysis using a ^{32}P-labelled human p53 cDNA probe. Polysomes used to calibrate the gradients were prepared in exactly the same way except for an additional purification step involving precipitation of the polysomes present in the homogenate with 100 mM MgCl$_2$ for 1 h on ice before sucrose gradient sedimentation. For calibration, 0.3-ml fractions were collected from the bottom of the gradient and A$_{254}$ of each fraction was determined.

Templates for in vitro transcription and translation

Plasmid p2516 contains nearly full-length human wild-type p53 cDNA and was constructed by the correct ligation of three cDNA fragments. One fragment corresponding to the 5' end of the p53 transcript was obtained from pR4-2 (Harlow *et al.*, 1985) after digestion with *Xba*I and *Pvu*II which cut in exons I and 5, respectively. The middle fragment was obtained from pProSp53 (Matlashewski *et al.*, 1987) after digestion with *Pvu*II and *Bam*HI which cut in exons 5 and 11, respectively. The third fragment corresponding to the 3' end of the p53 transcript was obtained by RT–PCR amplification of the 3'UTR of p53 mRNA using p53-specific oligonucleotides as primers, 3'SX13 (sense, exon 11, GTCACCCCATCCCACACCCTGG) and AS-4. The PCR-amplified fragment was end-filled with the Klenow fragment of DNA polymerase I and digested at an internal *Bam*HI site. These three fragments which represent contiguous sequences of the native p53 transcript were inserted between the *Xba*I and *Sma*I sites of a modified form of the pSP64 vector (Promega) in which polylinker sequences between the *Hin*dIII site and the *Xba*I site were deleted. The resulting plasmid is referred to as p2516 and yields a p53 transcript *in vitro* starting with the sequence 5'GAATACAAGCTCTAGA.....3'. The *in vitro* transcript is nearly identical to p53 transcripts originating from the most 3' of the major transcription initiation sites *in vivo* which start with 5'CAAAAGTCTAGA.....3' (Tuck and Crawford, 1989). The beginning of identity corresponding to an *Xba*I site in the cDNA is underlined. Digestion of p2516 with *Eco*RI provides a template that can produce a synthetic full-length p53 transcript of 2516 nucleotides. Digestion with *Bam*HI provides a template for a truncated p53 transcript of 2034 nucleotides that is missing sequences from the 3'UTR containing the Alu-like element.

The plasmid pCAT-Alu was constructed in two steps. First, the chloramphenicol acetyltransferase gene was excised from the CAT plasmid (Fu *et al.*, 1991) with *Hin*dIII and *Bam*HI, and inserted into pSP64 to generate pSP6CAT. Second, the *Bam*HI–*Eco*RI fragment from p2516 that contains the Alu-like element present in the p53 3'UTR was inserted immediately downstream of the CAT gene. The plasmid pCAT-BS was constructed by removing the *Sma*I–*Bam*HI fragment of the p53 3'UTR present in p2516 and inserting this fragment in reverse orientation into pSP6CAT immediately downstream of CAT. This *Sma*I–*Bam*HI fragment is missing the Alu-like element present at the distal end of the p53 3'UTR.

In vitro transcription and in vitro polyadenylation

Plasmid DNAs containing templates for *in vitro* transcription were linearized at selected restriction endonuclease sites. Standard transcription assays (Melton *et al.*, 1984) were performed as described above for the preparation of antisense RNA probes with the omission of [^{32}P]UTP. 0.5 mM ^{7m}G(5')ppp(5')G and 0.05 mM GTP were included in the reactions to provide efficient capping at the 5' end of synthetic transcripts. Polyadenylation reactions contained synthetic RNA, 0.2 mM ATP, 50 mM Tris–HCl pH 8.0, 10 mM MgCl$_2$, 250 mM NaCl, 2 mM MnCl$_2$, 2 mM dithiothreitol, 1 unit/μl RNAguard (Pharmacia), 500 μg/ml of BSA (Pharmacia) and 5 units of poly(A) polymerase (Pharmacia) in a 50 μl final volume (McGrew *et al.*, 1989). After 30 min at 37°C, polyadenylated RNAs were purified by phenol/chloroform extraction and ethanol precipitation.

In vitro translation and immunoprecipitation

Synthetic transcripts were translated in micrococcal-nuclease-treated rabbit reticulocyte lysates (Promega) under the conditions recommended by the supplier. Reactions containing p53 transcripts were incubated for 30 min at 30°C in the presence of [^{35}S]methionine and stopped by addition of dithiothreitol to a final concentration of 1 mM and EDTA pH 8.0 to a final concentration of 10 mM. Each reaction was then divided into two aliquots, one for immunoprecipitation with the p53-specific monoclonal antibody PAb421 and the other for immunoprecipitation with a control antibody PAb419. Reactions containing CAT or luciferase transcripts were incubated for 30 min at 30°C in the presence of [^{35}S]methionine and were stopped by addition of protein sample buffer, boiled for 5 min and resolved by polyacrylamide gel electrophoresis.

Acknowledgements

This work was supported by grants from the Medical Research Council of Canada and from the National Cancer Institute of Canada.

References

Banks,L., Matlashewski,G. and Crawford,L. (1986) Isolation of human p53-specific monoclonal antibodies and their use in the studies of human p53 expression. *Eur. J. Biochem.*, 159, 529–534.

Benchimol,S., Munroe,D.G., Peacock,J., Gray,D. and Smith,L.J. (1989) Abnormalities in structure and expression of the p53 gene in leukemia. *Cancer Cells*, 7, 121–125.

Bienz,B., Zakut-Houri,R., Givol,D. and Oren,M. (1984) Analysis of the gene coding for the murine cellular tumour antigen p53. *EMBO J.*, 3, 2179–2183.

Bienz-Tadmor,B., Zakut-Houri,R., Libresco,S., Givol,D. and Oren,M. (1985) The 5' region of the p53 gene: evolutionary conservation and evidence for a negative regulatory element. *EMBO J.*, 4, 3209–3213.

Chirgwin,J.M., Przybyla,A.E., McDonald,R.J. and Rutter,W.J. (1979) Isolation of biochemically active ribonucleic acid from sources enriched in ribonuclease. *Biochemistry*, 18, 5294–5299.

Evans,T.C., Crittenden,S.L., Kodoyianni,V. and Kimble,J. (1994) Translational control of material *glp-1* mRNA establishes an asymmetry in the *C. elegans* embryos. *Cell*, 77, 183–194.

Feinberg,A.P. and Vogelstein,B. (1983) A technique for radio-labeling DNA restriction endonuclease fragments to high specific activities. *Anal. Biochem.*, 132, 6–13.

Fenaux,P., Jonveaux,P., Quiquandon,I., Laï,J.L., Pignon,J.M., Loucheux-Lefebvre,M.H., Bauters,F., Berger,R. and Kerckaert,J.P. (1991) p53 gene mutations in acute myeloid leukemia with 17p monosomy. *Blood*, 78, 1652–1657.

Fenaux,P., Preudhomme,C., Quiquandon,I., Jonveaux,P., Laï,J.L., Vanrumbeke,M., Loucheux-Lefebvre,M.H., Bauters,F., Berger,R. and Kerckaert,J.P. (1992) Mutations of the p53 gene in acute myeloid leukaemia. *J. Haematol.*, 80, 178–183.

Feng,S. and Holland,E.C. (1988) HIV-1 *tat* trans-activation requires the loop sequence within *tat*. *Nature*, 334, 165–167.

Fialkow,P.J., Singer,J.W., Raskind,W.H., Adamson,J.W., Jacobson,R.J., Bernstein,I.D., Dow,L.W., Najfeld,V. and Veith,R. (1987) Clonal development, stem-cell differentiation, [illegible] nonlymphocytic leukemia. *N. Engl. J. Med.*. 317. 468–473.

Fritsche.M.. Haessler.C. and Brandner.G. (1993) Induction of nuclear accumulation of the tumor-suppressing protein p53 by DNA-damaging agents. *Oncogene*. 8. 307–318.

Fort.P.. Marty.L.. Piechaczyk.M.. Sabrouty.S.E.. Dani.C.. Jeanteur.P. and Blanchard.J.M. (1985) Various rat adult tissues express only one major mRNA species from the glyceraldehyde-3-phosphate-dehydrogenase multigenic family. *Nucleic Acids Res.*. 13. 1431–1442.

Fu.L.. Ye.R.. Browder.L.W. and Johnston.R.N. (1991) Translational potentiation of messenger RNA with secondary structure in *Xenopus*. *Science*. 251. 807–810.

Gavis.E.R. and Lehmann.R. (1994) Translational regulation of *nanos* by RNA localization. *Nature*. 369. 315–318.

Goodwin.E.B.. Okkema.P.O.. Evans.T.C. and Kimble.J. (1993) Translational regulation of *tra-2* by its 3′ UTR controls sexual identity in *C. elegans*. *Cell*. 75. 329–339.

Harlow.E.. Crawford.L.V.. Pim.D.C. and Williamson.N.M. (1981) Monoclonal antibodies specific for Simian virus 40 tumor antigen. *J. Virol.*. 39. 861–869.

Harlow.E.. Williamson.N.M.. Ralston.R.. Helfman.D.M. and Adams.T.E. (1985) Molecular cloning and *in vitro* expression of a DNA clone for human cellular tumor antigen p53. *Mol. Cell. Biol.*. 5. 1601–1610.

Herzog.H.. Hofferer.L.. Schneider.R. and Schweiger.M. (1990) cDNA encoding the human homologue of rat ribosomal protein L35a. *Nucleic Acids Res.*. 18. 4600.

Hsu.H.C.. Tseng.H.J.. Lai.P.L.. Lee.P.H. and Peng.S.Y. (1993) Expression of p53 gene in 184 unifocal hepatocellular carcinomas: association with tumor growth and invasiveness. *Cancer Res.*. 53. 4691–4694.

Huarte.J.. Stutz.A.. O'Connell.M.L.. Gubler.P.. Belin.D.. Darrow.A.L.. Strickland.S. and Vassalli.J.D. (1992) Transient translational silencing by reversible mRNA deadenylation. *Cell*. 69. 1021–1030.

Jackson.R.J. (1993) Cytoplasmic regulation of mRNA function: the importance of the 3′ UTR. *Cell*. 74. 9–14.

Kastan.M.B. *et al.* (1991a) Levels of p53 protein increase with maturation in human hematopoietic cells. *Cancer Res.*. 51. 4279–4286.

Kastan.M.B.. Onyekwere.O.. Sidransky.D.. Vogelstein.B. and Craig.R.W. (1991b) Participation of p53 protein in the cellular response to DNA damage. *Cancer Res.*. 51. 6304–6311.

Kupiec.J.J.. Giron.M.L.. Vilette.D.. Jeltsch.J.M. and Emanoil-Ravier.R. (1987) Isolation of high-molecular-weight DNA from eukaryotic cells by formamide treatment and dialysis. *Anal. Biochem.*. 164. 53–59.

Kwon.Y.K. and Hecht.N.B. (1993) Binding of a phosphoprotein to the 3′ untranslated region of the mouse protamine 2 mRNA temporally represses its translation. *Mol. Cell. Biol.*. 13. 6547–6557.

Lai.J.L.. Preudhomme.C.. Zandecki.M.. Flactif.M.. Vanrumbeke.M.. Lepelley.P.. Wattel.E. and Fenaux.P. (1995) Myelodysplastic syndromes and acute myeloid leukemia with 17p deletion. An entity characterized by specific dysgranulopoiesis and a high incidence of P53 mutations. *Leukemia*. 9. 370–381.

Lamb.P. and Crawford.L. (1986) Characterization of the human p53 gene. *Mol. Cell. Biol.*. 6. 1379–1385.

Lu.X. and Lane.D.P. (1993) Differential induction of transcriptionally active p53 following UV or ionizing radiation: Defects in chromosome instability syndromes? *Cell*. 75. 765–778.

Matlashewski.G.. Lamb.P.. Pim.D.. Peacock.J.. Crawford.L. and Benchimol.S. (1984) Isolation and characterization of a human p53 cDNA clone: expression of the human p53 gene. *EMBO J.*. 13. 3257–3262.

Matlashewski.G.. Banks.L.. Pim.D. and Crawford.L.V. (1986) Analysis of human p53 proteins and mRNA levels in normal and transformed cells. *Eur. J. Biochem.*. 154. 665–672.

Matlashewski.G.J.. Tuck.S.. Pim.D.. Lamb.P.. Schneider.J. and Crawford. L.V. (1987) Primary structure polymorphism at amino residue 72 of human p53. *Mol. Cell. Biol.*. 7. 961–963.

Melefors.O. and Hentze.M.W. (1993) Translational regulation by mRNA–protein interactions in eukaryotic cells: ferritin and beyond. *BioEssays*. 15. 85–90.

Melton.D.A.. Krieg.P.. Rebagliati.M.. Maniatis.T.. Zinn.K. and Green.M. (1984) Efficient *in vitro* synthesis of biologically active RNA and RNA hybridization probes from plasmids containing a bacteriophage SP6 promoter. *Nucleic Acids Res.*. 12. 7035–7056.

Milner.J. (1991) A conformation hypothesis for the suppressor and promoter functions of p53 in cell growth control and in cancer. *Proc. R. Soc. Lond. B.*. 245. 139–145.

blast cell and T-lymphocyte progenitors in the blood of patients with acute myeloblastic leukemia. *Blood*. 54. 186–195.

Moll.U.M.. Riou.G. and Levine.A.J. (1992) Two distinct mechanisms alter p53 in breast cancer: mutation and nuclear exclusion. *Proc. Natl Acad. Sci. USA*. 89. 7262–7266.

Moll.U.M.. LaQuaglia.M.. Benard.J. and Riou.G. (1995) Wild-type p53 protein undergoes cytoplasmic sequestration in undifferentiated neuroblastomas but not in differentiated tumors. *Proc. Natl Acad. Sci. USA*. 92. 4407–4411.

Momand.J.. Zambetti.G.P.. Olson.D.C.. George.D. and Levine.A.J. (1992) The *mdm-2* oncogene product forms a complex with the p53 protein and inhibits p53-mediated transactivation. *Cell*. 69. 1237–1245.

Mosner.J.. Mummenbrauer.T.. Bauer.C.. Sczakiel.G.. Grosse.F. and Deppert.W. (1995) Negative feedback regulation of wild-type p53 biosynthesis. *EMBO J.*. 12. 4739–4746.

Oliner.J.D.. Kinzler.K.W.. Meltzer.P.S.. George.D and Vogelstein.B. (1992) Amplification of a gene encoding a p53-associated protein in human sarcomas. *Nature*. 358. 80–83.

Papayannopoulou.T.. Nakomoto.B.. Kurachi.S.. Tweeddale.M. and Messner.H. (1988) Surface antigenic profile and globin phenotype of two new human erythroleukemia lines: characterization and interpretations. *Blood*. 72. 1029–1038.

Pause.A.. Methot.N. and Sonenberg.N. (1993) The HRIGRXXR region of the DEAD box RNA helicase eukaryotic translation initiation factor 4A is required for RNA binding and ATP hydrolysis. *Mol. Cell. Biol.*. 13. 6789–6798.

Rogel.A.. Popliker.M.. Webb.C.G. and Oren.M. (1985) p53 cellular tumor antigen: analysis of mRNA levels in normal adult tissues. embryos. and tumors. *Mol. Cell. Biol.*. 5. 2851–2855.

Sasano.H.. Goukon.Y.. Nishihia.T. and Nagura.H. (1992) *In situ* hybridization and immunohistochemistry of p53 tumor suppressor gene in human esophageal carcinoma. *Am. J. Pathol.*. 141. 545–550.

Scheffner.M.. Werness.B.A.. Huibregtse.J.M.. Levine.A.J. and Howley. P.M. (1990) The E6 oncoprotein encoded by human papillomavirus types 16 and 18 promotes the degradation of p53. *Cell*. 63. 1129–1136.

Slingerland.J.M.. Minden.M.D. and Benchimol.S. (1991) Mutations of the p53 gene in human acute myelogenous leukemia. *Blood*. 7. 1500–1507.

Smith.L.J.. McCulloch.E.A. and Benchimol.S. (1986) Expression of the p53 oncogene in acute myeloblastic leukemia. *J. Exp. Med.*. 164. 751–761.

Sugimoto.K.. Toyoshima.H.. Sakai.R.. Miyagawa.K.. Hagiwara.K.. Hirai.H.. Ishikana.F. and Takaku.F. (1991) Mutations of the p53 gene in lymphoid leukemia. *Blood*. 77. 1153–1156.

Sugimoto.K.. Hirano.N.. Toyoshima.H.. Chiba.S.. Mano.H.. Takaku.F.. Yazaki.Y. and Hirai.H. (1993) Mutations of the p53 gene in myelodysplastic syndromes and MDS-derived leukemia. *Blood*. 81. 3022–3026.

Torczynski.R.M.. Fuke.M. and Bollon.A.P. (1985) Cloning and sequencing of a human 18S ribosomal RNA gene. *DNA*. 4. 283–291.

Trecca.D.. Longo.L.. Biondi.A.. Cro.L.. Calori.R.. Grignani.F.. Maiolo.A.T.. Pelicci.P.G. and Neri.A. (1994) Analysis of p53 gene mutations in acute myeloid leukemia. *Am. J. Hematol.*. 46. 304–309.

Tuck.S.P. and Crawford.L. (1989) Characterization of the human p53 gene promoter. *Mol. Cell. Biol.*. 9. 2163–2172.

Ueda.H.. Ullrich.S.J.. Gangemi.J.D.. Kappel.C.A.. Ngo.L.. Feitelson. M.A. and Jay.G. (1995) Functional inactivation but not structural mutation of p53 causes liver cancer. *Nature Genet.*. 9. 41–47.

Ullrich.S.J.. Mercer.W.E. and Appella.E. (1992) Human wild-type p53 adopts a unique conformational and phosphorylation state *in vivo* during growth arrest of glioblastoma cells. *Oncogene*. 7. 1635–1643.

Wang.C.. Curtis.J.E.. Minden.M.D. and McCulloch.E.A. (1989) Expression of a retinoic acid receptor gene in myeloid leukemia cells. *Leukemia*. 3. 264–269.

Wattel.E.. Preudhomme.C.. Hecquet.B.. Vanrumbeke.M.. Quesnel.B.. Dervite.I.. Morel.P. and Fenaux.P. (1994) p53 mutations are associated with resistance to chemotherapy and short survival in hematologic malignancies. *Blood*. 84. 3148–3157.

Wiggans.R.G.. Jacobson.R.J.. Fialkow.P.J.. Woolley.P.V.. Macdonald.S.J. and Schein.P.S. (1978) Probable clonal origin of acute myeloblastic leukemia following radiation and chemotherapy of colon cancer. *Blood*. 52. 650–663.

Winship.P.R. (1989) An improved method for directly sequencing PCR amplified material using dimethyl sulphoxide. *Nucleic Acids Res.*. 17. 1266.

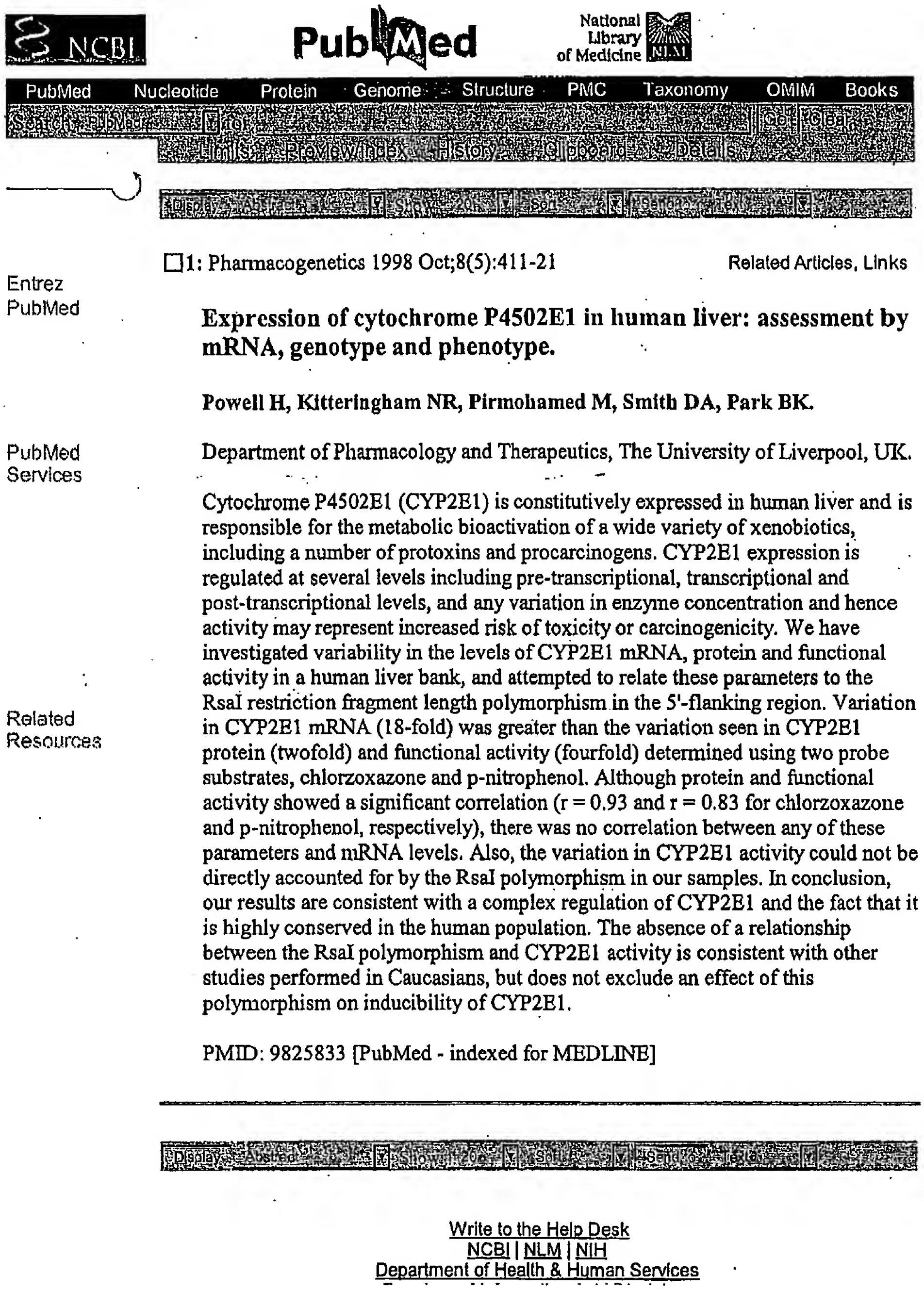
Winter,E., Yamamoto,F., Almoguera,C. [illegible] to detect and characterize point mutations in transcribed genes: amplification and overexpression of the mutant C-Ki-ras allele in human tumor cells. *Proc. Natl Acad. Sci. USA*, 82, 7575–7579.

Zhan,Q., Carrier,F. and Fornace,A.J.,Jr (1993) Induction of cellular p53 activity by DNA-damaging agents and growth arrest. *Mol. Cell. Biol.*, 13, 4242–4250.

Zhang,W., Hu,G., Estey,E., Hester,J. and Deisseroth,A. (1992) Altered conformation of the p53 protein in myeloid leukemia cells and mitogen-stimulated normal blood cells. *Oncogene*, 7, 1645–1647.

Zhu,Y.M., Bradbury,D. and Russell,N. (1993) Expression of different conformations of p53 in the blast cells of acute myeloblastic leukaemia is related to *in vitro* growth characteristics. *Br. J. Cancer*, 68, 851–855.

Received on November 30, 1995; revised on April 18, 1996



□1: Pharmacogenetics 1998 Oct;8(5):411-21

Expression of cytochrome P4502E1 in human liver: assessment by mRNA, genotype and phenotype.

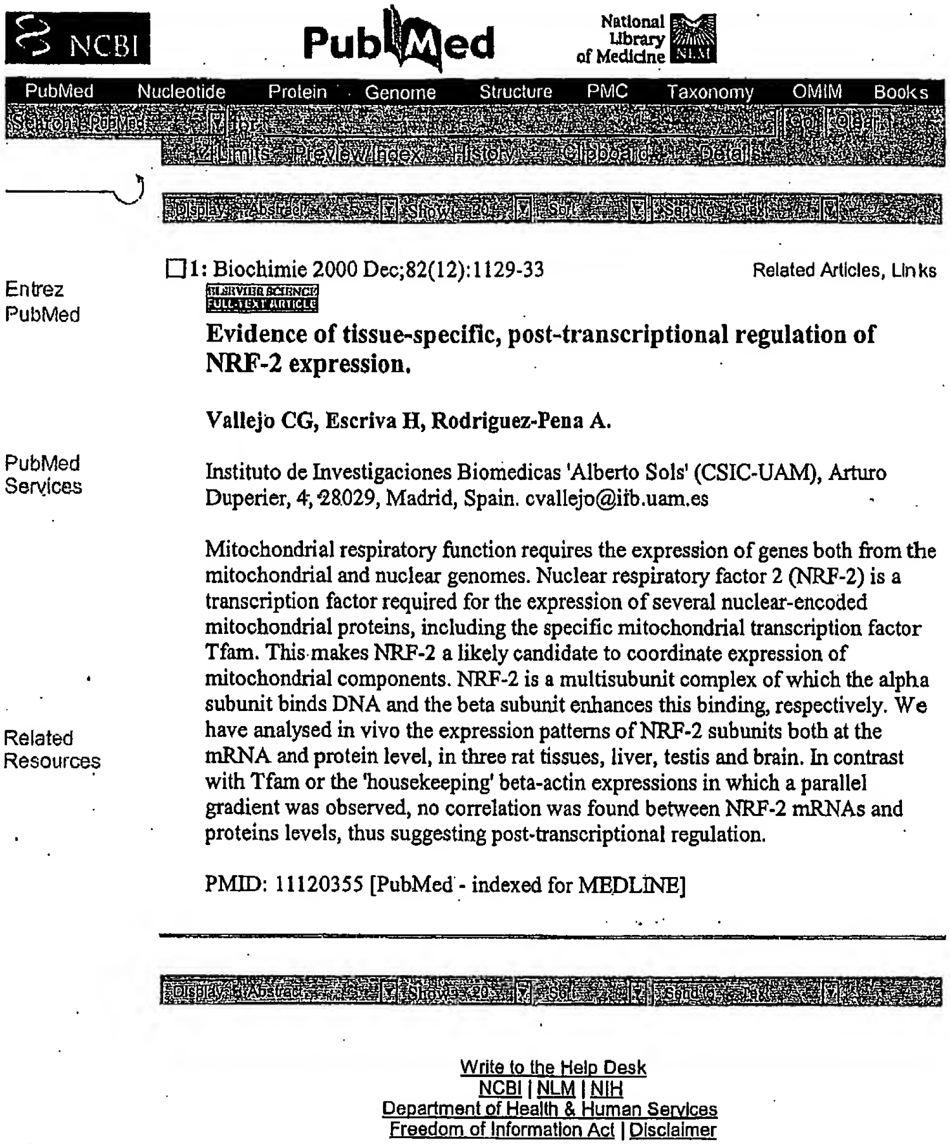
Powell H, Kitteringham NR, Pirmohamed M, Smith DA, Park BK.

Department of Pharmacology and Therapeutics, The University of Liverpool, UK.

Cytochrome P4502E1 (CYP2E1) is constitutively expressed in human liver and is responsible for the metabolic bioactivation of a wide variety of xenobiotics, including a number of protoxins and procarcinogens. CYP2E1 expression is regulated at several levels including pre-transcriptional, transcriptional and post-transcriptional levels, and any variation in enzyme concentration and hence activity may represent increased risk of toxicity or carcinogenicity. We have investigated variability in the levels of CYP2E1 mRNA, protein and functional activity in a human liver bank, and attempted to relate these parameters to the RsaI restriction fragment length polymorphism in the 5'-flanking region. Variation in CYP2E1 mRNA (18-fold) was greater than the variation seen in CYP2E1 protein (twofold) and functional activity (fourfold) determined using two probe substrates, chlorzoxazone and p-nitrophenol. Although protein and functional activity showed a significant correlation ($r = 0.93$ and $r = 0.83$ for chlorzoxazone and p-nitrophenol, respectively), there was no correlation between any of these parameters and mRNA levels. Also, the variation in CYP2E1 activity could not be directly accounted for by the RsaI polymorphism in our samples. In conclusion, our results are consistent with a complex regulation of CYP2E1 and the fact that it is highly conserved in the human population. The absence of a relationship between the RsaI polymorphism and CYP2E1 activity is consistent with other studies performed in Caucasians, but does not exclude an effect of this polymorphism on inducibility of CYP2E1.

PMID: 9825833 [PubMed - indexed for MEDLINE]

Write to the Help Desk
NCBI | NLM | NIH
Department of Health & Human Services



□1: Biochimie 2000 Dec;82(12):1129-33

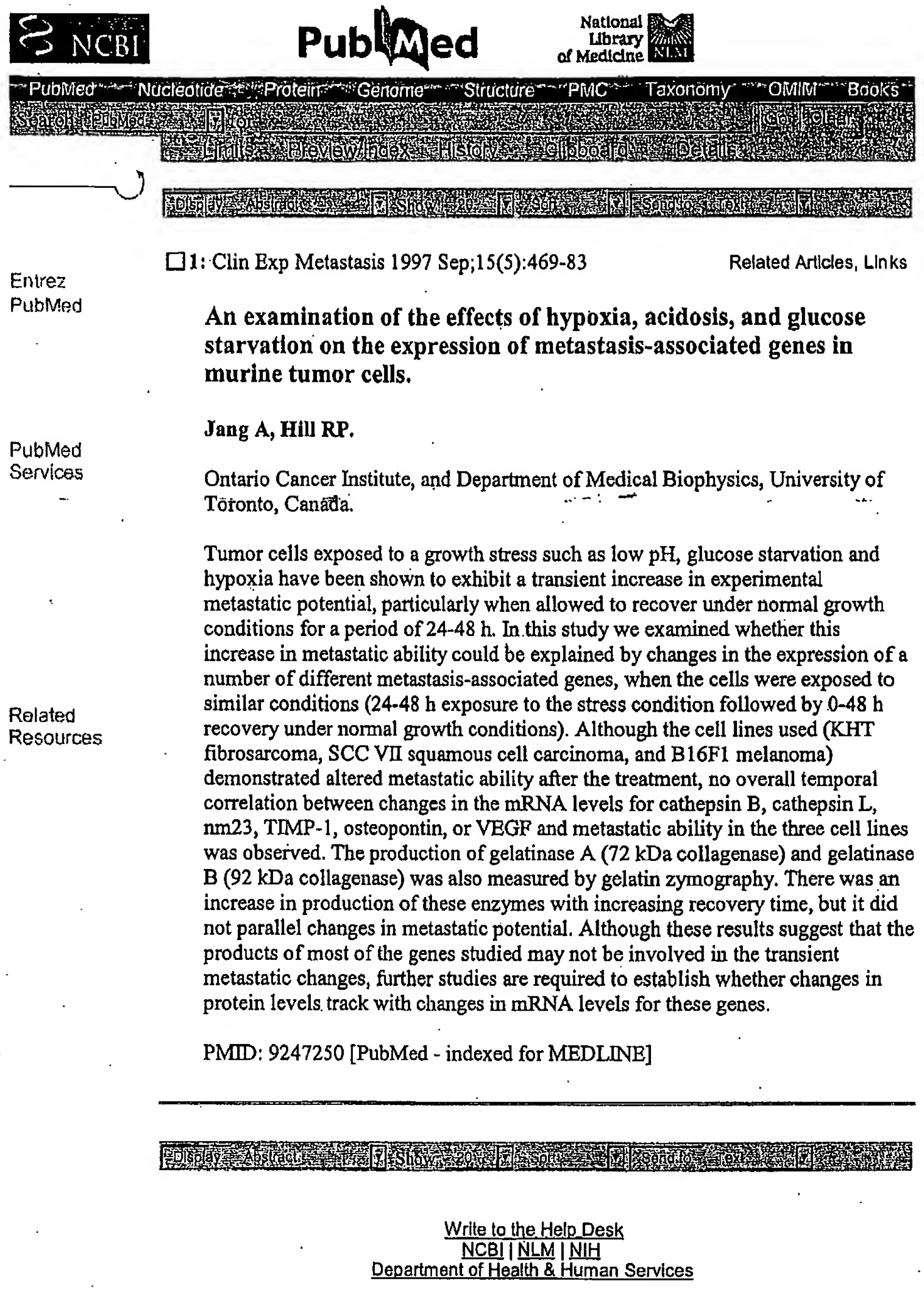
Evidence of tissue-specific, post-transcriptional regulation of NRF-2 expression.

Vallejo CG, Escriva H, Rodriguez-Pena A.

Instituto de Investigaciones Biomedicas 'Alberto Sols' (CSIC-UAM), Arturo Duperier, 4, 28029, Madrid, Spain. cvallejo@iib.uam.es

Mitochondrial respiratory function requires the expression of genes both from the mitochondrial and nuclear genomes. Nuclear respiratory factor 2 (NRF-2) is a transcription factor required for the expression of several nuclear-encoded mitochondrial proteins, including the specific mitochondrial transcription factor Tfam. This makes NRF-2 a likely candidate to coordinate expression of mitochondrial components. NRF-2 is a multisubunit complex of which the alpha subunit binds DNA and the beta subunit enhances this binding, respectively. We have analysed in vivo the expression patterns of NRF-2 subunits both at the mRNA and protein level, in three rat tissues, liver, testis and brain. In contrast with Tfam or the 'housekeeping' beta-actin expressions in which a parallel gradient was observed, no correlation was found between NRF-2 mRNAs and proteins levels, thus suggesting post-transcriptional regulation.

PMID: 11120355 [PubMed - indexed for MEDLINE]


NCBI
PubMed
National Library of Medicine NLM
PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Books
Search PubMed for Go Clear
Limits Preview/Index History Clipboard Details
Display Abstract Show 20 Sort Send to Text
Entrez PubMed
PubMed Services
Related Resources
Related Articles, Links

□ 1: Clin Exp Metastasis 1997 Sep;15(5):469-83

An examination of the effects of hypoxia, acidosis, and glucose starvation on the expression of metastasis-associated genes in murine tumor cells.

Jang A, Hill RP.

Ontario Cancer Institute, and Department of Medical Biophysics, University of Toronto, Canada.

Tumor cells exposed to a growth stress such as low pH, glucose starvation and hypoxia have been shown to exhibit a transient increase in experimental metastatic potential, particularly when allowed to recover under normal growth conditions for a period of 24-48 h. In this study we examined whether this increase in metastatic ability could be explained by changes in the expression of a number of different metastasis-associated genes, when the cells were exposed to similar conditions (24-48 h exposure to the stress condition followed by 0-48 h recovery under normal growth conditions). Although the cell lines used (KHT fibrosarcoma, SCC VII squamous cell carcinoma, and B16F1 melanoma) demonstrated altered metastatic ability after the treatment, no overall temporal correlation between changes in the mRNA levels for cathepsin B, cathepsin L, nm23, TIMP-1, osteopontin, or VEGF and metastatic ability in the three cell lines was observed. The production of gelatinase A (72 kDa collagenase) and gelatinase B (92 kDa collagenase) was also measured by gelatin zymography. There was an increase in production of these enzymes with increasing recovery time, but it did not parallel changes in metastatic potential. Although these results suggest that the products of most of the genes studied may not be involved in the transient metastatic changes, further studies are required to establish whether changes in protein levels track with changes in mRNA levels for these genes.

PMID: 9247250 [PubMed - indexed for MEDLINE]

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 14717–14722, December 1998
Cell Biology, Medical Sciences

WISP genes are members of the connective tissue growth factor family that are up-regulated in Wnt-1-transformed cells and aberrantly expressed in human colon tumors

DIANE PENNICA*†, TODD A. SWANSON*, JAMES W. WELSH*, MARGARET A. ROY‡, DAVID A. LAWRENCE*, JAMES LEE‡, JENNIFER BRUSH‡, LISA A. TANEYHILL§, BETHANNE DEUEL‡, MICHAEL LEW¶, COLIN WATANABE||, ROBERT L. COHEN*, MONA F. MELHEM**, GENE G. FINLEY**, PHIL QUIRKE††, AUDREY D. GODDARD‡, KENNETH J. HILLAN¶, AUSTIN L. GURNEY‡, DAVID BOTSTEIN‡,‡‡, AND ARNOLD J. LEVINE§

Departments of *Molecular Oncology, ‡Molecular Biology, ||Scientific Computing, and ¶Pathology, Genentech Inc., 1 DNA Way, South San Francisco, CA 94080; **University of Pittsburgh School of Medicine, Veterans Administration Medical Center, Pittsburgh, PA 15240; ††University of Leeds, Leeds, LS29JT United Kingdom; ‡‡Department of Genetics, Stanford University, Palo Alto, CA 94305; and §Department of Molecular Biology, Princeton University, Princeton, NJ 08544

Contributed by David Botstein and Arnold J. Levine, October 21, 1998

ABSTRACT Wnt family members are critical to many developmental processes, and components of the Wnt signaling pathway have been linked to tumorigenesis in familial and sporadic colon carcinomas. Here we report the identification of two genes, *WISP-1* and *WISP-2*, that are up-regulated in the mouse mammary epithelial cell line C57MG transformed by Wnt-1, but not by Wnt-4. Together with a third related gene, *WISP-3*, these proteins define a subfamily of the connective tissue growth factor family. Two distinct systems demonstrated *WISP* induction to be associated with the expression of Wnt-1. These included (*i*) C57MG cells infected with a Wnt-1 retroviral vector or expressing Wnt-1 under the control of a tetracyline repressible promoter, and (*ii*) Wnt-1 transgenic mice. The *WISP-1* gene was localized to human chromosome 8q24.1–8q24.3. *WISP-1* genomic DNA was amplified in colon cancer cell lines and in human colon tumors and its RNA overexpressed (2- to >30-fold) in 84% of the tumors examined compared with patient-matched normal mucosa. *WISP-3* mapped to chromosome 6q22–6q23 and also was overexpressed (4- to >40-fold) in 63% of the colon tumors analyzed. In contrast, *WISP-2* mapped to human chromosome 20q12–20q13 and its DNA was amplified, but RNA expression was reduced (2- to >30-fold) in 79% of the tumors. These results suggest that the *WISP* genes may be downstream of Wnt-1 signaling and that aberrant levels of *WISP* expression in colon cancer may play a role in colon tumorigenesis.

Wnt-1 is a member of an expanding family of cysteine-rich, glycosylated signaling proteins that mediate diverse developmental processes such as the control of cell proliferation, adhesion, cell polarity, and the establishment of cell fates (1, 2). Wnt-1 originally was identified as an oncogene activated by the insertion of mouse mammary tumor virus in virus-induced mammary adenocarcinomas (3, 4). Although Wnt-1 is not expressed in the normal mammary gland, expression of Wnt-1 in transgenic mice causes mammary tumors (5).

In mammalian cells, Wnt family members initiate signaling by binding to the seven-transmembrane spanning Frizzled receptors and recruiting the cytoplasmic protein Dishevelled (Dsh) to the cell membrane (1, 2, 6). Dsh then inhibits the kinase activity of the normally constitutively active glycogen synthase kinase-3β (GSK-3β) resulting in an increase in β-catenin levels. Stabilized β-catenin interacts with the transcription factor TCF/Lef1, forming a complex that appears in the nucleus and binds TCF/Lef1 target DNA elements to activate transcription (7, 8). Other experiments suggest that the adenomatous polyposis coli (APC) tumor suppressor gene also plays an important role in Wnt signaling by regulating β-catenin levels (9). APC is phosphorylated by GSK-3β, binds to β-catenin, and facilitates its degradation. Mutations in either APC or β-catenin have been associated with colon carcinomas and melanomas, suggesting these mutations contribute to the development of these types of cancer, implicating the Wnt pathway in tumorigenesis (1).

Although much has been learned about the Wnt signaling pathway over the past several years, only a few of the transcriptionally activated downstream components activated by Wnt have been characterized. Those that have been described cannot account for all of the diverse functions attributed to Wnt signaling. Among the candidate Wnt target genes are those encoding the nodal-related 3 gene, *Xnr3*, a member of the transforming growth factor (TGF)-β superfamily, and the homeobox genes, *engrailed*, *goosecoid*, *twin* (*Xtwn*), and *siamois* (2). A recent report also identifies c-*myc* as a target gene of the Wnt signaling pathway (10).

To identify additional downstream genes in the Wnt signaling pathway that are relevant to the transformed cell phenotype, we used a PCR-based cDNA subtraction strategy, suppression subtractive hybridization (SSH) (11), using RNA isolated from C57MG mouse mammary epithelial cells and C57MG cells stably transformed by a Wnt-1 retrovirus. Overexpression of Wnt-1 in this cell line is sufficient to induce a partially transformed phenotype, characterized by elongated and refractile cells that lose contact inhibition and form a multilayered array (12, 13). We reasoned that genes differentially expressed between these two cell lines might contribute to the transformed phenotype.

In this paper, we describe the cloning and characterization of two genes up-regulated in Wnt-1 transformed cells, *WISP-1* and *WISP-2*, and a third related gene, *WISP-3*. The *WISP* genes are members of the CCN family of growth factors, which includes connective tissue growth factor (CTGF), Cyr61, and *nov*, a family not previously linked to Wnt signaling.

MATERIALS AND METHODS

SSH. SSH was performed by using the PCR-Select cDNA Subtraction Kit (CLONTECH). Tester double-stranded

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

PNAS is available online at www.pnas.org.

Abbreviations: TGF, transforming growth factor; CTGF, connective tissue growth factor; SSH, suppression subtractive hybridization; VWC, von Willebrand factor type C module.

Data deposition: The sequences reported in this paper have been deposited in the Genbank database (accession nos. AF100777, AF100778, AF100779, AF100780, and AF100781).

†To whom reprint requests should be addressed. e-mail: diane@gene.com.

cDNA was synthesized from 2 μg of poly(A)$^+$ RNA isolated from the C57MG/Wnt-1 cell line and driver cDNA from 2 μg of poly(A)$^+$ RNA from the parent C57MG cells. The subtracted cDNA library was subcloned into a pGEM-T vector for further analysis.

cDNA Library Screening. Clones encoding full-length mouse *WISP-1* were isolated by screening a λgt10 mouse embryo cDNA library (CLONTECH) with a 70-bp probe from the original partial clone 568 sequence corresponding to amino acids 128–169. Clones encoding full-length human *WISP-1* were isolated by screening λgt10 lung and fetal kidney cDNA libraries with the same probe at low stringency. Clones encoding full-length mouse and human *WISP-2* were isolated by screening a C57MG/Wnt-1 or human fetal lung cDNA library with a probe corresponding to nucleotides 1463–1512. Full-length cDNAs encoding *WISP-3* were cloned from human bone marrow and fetal kidney libraries.

Expression of Human *WISP* RNA. PCR amplification of first-strand cDNA was performed with human Multiple Tissue cDNA panels (CLONTECH) and 300 μM of each dNTP at 94°C for 1 sec, 62°C for 30 sec, 72°C for 1 min, for 22–32 cycles. *WISP* and glyceraldehyde-3-phosphate dehydrogenase primer sequences are available on request.

***In Situ* Hybridization.** ^{33}P-labeled sense and antisense riboprobes were transcribed from an 897-bp PCR product corresponding to nucleotides 601–1440 of mouse *WISP-1* or a 294-bp PCR product corresponding to nucleotides 82–375 of mouse *WISP-2*. All tissues were processed as described (40).

Radiation Hybrid Mapping. Genomic DNA from each hybrid in the Stanford G3 and Genebridge4 Radiation Hybrid Panels (Research Genetics, Huntsville, AL) and human and hamster control DNAs were PCR-amplified, and the results were submitted to the Stanford or Massachusetts Institute of Technology web servers.

Cell Lines, Tumors, and Mucosa Specimens. Tissue specimens were obtained from the Department of Pathology (University of Pittsburgh) for patients undergoing colon resection and from the University of Leeds, United Kingdom. Genomic DNA was isolated (Qiagen) from the pooled blood of 10 normal human donors, surgical specimens, and the following ATCC human cell lines: SW480, COLO 320DM, HT-29, WiDr, and SW403 (colon adenocarcinomas), SW620 (lymph node metastasis, colon adenocarcinoma), HCT 116 (colon carcinoma), SK-CO-1 (colon adenocarcinoma, ascites), and HM7 (a variant of ATCC colon adenocarcinoma cell line LS 174T). DNA concentration was determined by using Hoechst dye 33258 intercalation fluorimetry. Total RNA was prepared by homogenization in 7 M GuSCN followed by centrifugation over CsCl cushions or prepared by using RNAzol.

Gene Amplification and RNA Expression Analysis. Relative gene amplification and RNA expression of *WISPs* and c-*myc* in the cell lines, colorectal tumors, and normal mucosa were determined by quantitative PCR. Gene-specific primers and fluorogenic probes (sequences available on request) were designed and used to amplify and quantitate the genes. The relative gene copy number was derived by using the formula $2^{(\Delta Ct)}$ where ΔCt represents the difference in amplification cycles required to detect the *WISP* genes in peripheral blood lymphocyte DNA compared with colon tumor DNA or colon tumor RNA compared with normal mucosal RNA. The ∂-method was used for calculation of the SE of the gene copy number or RNA expression level. The *WISP*-specific signal was normalized to that of the glyceraldehyde-3-phosphate dehydrogenase housekeeping gene. All TaqMan assay reagents were obtained from Perkin–Elmer Applied Biosystems.

RESULTS

Isolation of *WISP-1* and *WISP-2* by SSH. To identify Wnt-1-inducible genes, we used the technique of SSH using the mouse mammary epithelial cell line C57MG and C57MG cells that stably express Wnt-1 (11). Candidate differentially expressed cDNAs (1,384 total) were sequenced. Thirty-nine percent of the sequences matched known genes or homologues, 32% matched expressed sequence tags, and 29% had no match. To confirm that the transcript was differentially expressed, semiquantitative reverse transcription–PCR and Northern analysis were performed by using mRNA from the C57MG and C57MG/Wnt-1 cells.

Two of the cDNAs, *WISP-1* and *WISP-2*, were differentially expressed, being induced in the C57MG/Wnt-1 cell line, but not in the parent C57MG cells or C57MG cells overexpressing Wnt-4 (Fig. 1 *A* and *B*). Wnt-4, unlike Wnt-1, does not induce the morphological transformation of C57MG cells and has no effect on β-catenin levels (13, 14). Expression of *WISP-1* was up-regulated approximately 3-fold in the C57MG/Wnt-1 cell line and *WISP-2* by approximately 5-fold by both Northern analysis and reverse transcription–PCR.

An independent, but similar, system was used to examine *WISP* expression after Wnt-1 induction. C57MG cells expressing the *Wnt-1* gene under the control of a tetracycline-repressible promoter produce low amounts of Wnt-1 in the repressed state but show a strong induction of *Wnt-1* mRNA and protein within 24 hr after tetracycline removal (8). The levels of Wnt-1 and *WISP* RNA isolated from these cells at various times after tetracycline removal were assessed by quantitative PCR. Strong induction of Wnt-1 mRNA was seen as early as 10 hr after tetracycline removal. Induction of *WISP* mRNA (2- to 6-fold) was seen at 48 and 72 hr (data not shown). These data support our previous observations that show that *WISP* induction is correlated with Wnt-1 expression. Because the induction is slow, occurring after approximately 48 hr, the induction of *WISPs* may be an indirect response to Wnt-1 signaling.

cDNA clones of human *WISP-1* were isolated and the sequence compared with mouse *WISP-1*. The cDNA sequences of mouse and human *WISP-1* were 1,766 and 2,830 bp in length, respectively, and encode proteins of 367 aa, with predicted relative molecular masses of ≈40,000 (M_r 40 K). Both have hydrophobic N-terminal signal sequences, 38 conserved cysteine residues, and four potential N-linked glycosylation sites and are 84% identical (Fig. 2*A*).

Full-length cDNA clones of mouse and human *WISP-2* were 1,734 and 1,293 bp in length, respectively, and encode proteins of 251 and 250 aa, respectively, with predicted relative molecular masses of ≈27,000 (M_r 27 K) (Fig. 2*B*). Mouse and human *WISP-2* are 73% identical. Human *WISP-2* has no potential N-linked glycosylation sites, and mouse *WISP-2* has one at

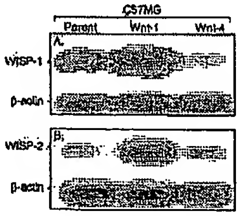


FIG. 1. *WISP-1* and *WISP-2* are induced by Wnt-1, but not Wnt-4, expression in C57MG cells. Northern analysis of *WISP-1* (*A*) and *WISP-2* (*B*) expression in C57MG, C57MG/Wnt-1, and C57MG/Wnt-4 cells. Poly(A)$^+$ RNA (2 μg) was subjected to Northern blot analysis and hybridized with a 70-bp mouse *WISP-1*-specific probe (amino acids 278–300) or a 190-bp *WISP-2*-specific probe (nucleotides 1438–1627) in the 3′ untranslated region. Blots were rehybridized with human β-actin probe.

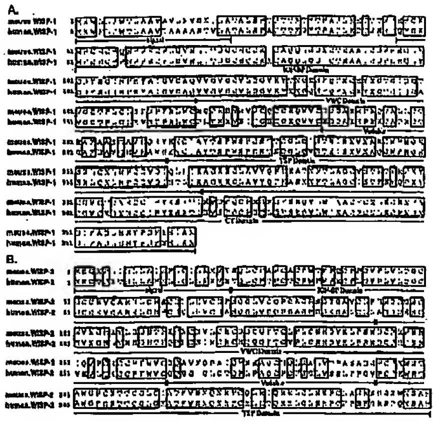


Fig. 2. Encoded amino acid sequence alignment of mouse and human *WISP-1* (*A*) and mouse and human *WISP-2* (*B*). The potential signal sequence, insulin-like growth factor-binding protein (IGF-BP), VWC, thrombospondin (TSP), and C-terminal (CT) domains are underlined.

position 197. *WISP-2* has 28 cysteine residues that are conserved among the 38 cysteines found in *WISP-1*.

Identification of *WISP-3*. To search for related proteins, we screened expressed sequence tag (EST) databases with the WISP-1 protein sequence and identified several ESTs as potentially related sequences. We identified a homologous protein that we have called WISP-3. A full-length human *WISP-3* cDNA of 1,371 bp was isolated corresponding to those ESTs that encode a 354-aa protein with a predicted molecular mass of 39,293. WISP-3 has two potential N-linked glycosylation sites and 36 cysteine residues. An alignment of the three human WISP proteins shows that WISP-1 and WISP-3 are the most similar (42% identity), whereas WISP-2 has 37% identity with WISP-1 and 32% identity with WISP-3 (Fig. 3*A*).

***WISPs* Are Homologous to the CTGF Family of Proteins.** Human *WISP-1*, *WISP-2*, and *WISP-3* are novel sequences; however, mouse *WISP-1* is the same as the recently identified *Elm1* gene. *Elm1* is expressed in low, but not high, metastatic mouse melanoma cells, and suppresses the *in vivo* growth and metastatic potential of K-1735 mouse melanoma cells (15). Human and mouse *WISP-2* are homologous to the recently described rat gene, *rCop-1* (16). Significant homology (36–44%) was seen to the CCN family of growth factors. This family includes three members, CTGF, Cyr61, and the protooncogene *nov*. CTGF is a chemotactic and mitogenic factor for fibroblasts that is implicated in wound healing and fibrotic disorders and is induced by TGF-β (17). Cyr61 is an extracellular matrix signaling molecule that promotes cell adhesion, proliferation, migration, angiogenesis, and tumor growth (18, 19). *nov* (nephroblastoma overexpressed) is an immediate early gene associated with quiescence and found altered in Wilms tumors (20). The proteins of the CCN family share functional, but not sequence, similarity to Wnt-1. All are secreted, cysteine-rich heparin binding glycoproteins that associate with the cell surface and extracellular matrix.

WISP proteins exhibit the modular architecture of the CCN family, characterized by four conserved cysteine-rich domains (Fig. 3*B*) (21). The N-terminal domain, which includes the first 12 cysteine residues, contains a consensus sequence (GCGCCXXC) conserved in most insulin-like growth factor (IGF)-binding proteins (BP). This sequence is conserved in WISP-2 and WISP-3, whereas WISP-1 has a glutamine in the third position instead of a glycine. CTGF recently has been shown to specifically bind IGF (22) and a truncated *nov* protein lacking the IGF-BP domain is oncogenic (23). The von Willebrand factor type C module (VWC), also found in certain collagens and mucins, covers the next 10 cysteine residues, and is thought to participate in protein complex formation and oligomerization (24). The VWC domain of WISP-3 differs from all CCN family members described previously, in that it contains only six of the 10 cysteine residues (Fig. 3 *A* and *B*). A short variable region follows the VWC domain. The third module, the thrombospondin (TSP) domain is involved in binding to sulfated glycoconjugates and contains six cysteine residues and a conserved WSxCSxxCG motif first identified in thrombospondin (25). The C-terminal (CT) module containing the remaining 10 cysteines is thought to be involved in dimerization and receptor binding (26). The CT domain is present in all CCN family members described to date but is absent in WISP-2 (Fig. 3 *A* and *B*). The existence of a putative signal sequence and the absence of a transmembrane domain suggest that WISPs are secreted proteins, an observation supported by an analysis of their expression and secretion from mammalian cell and baculovirus cultures (data not shown).

Expression of *WISP* mRNA in Human Tissues. Tissue-specific expression of human *WISPs* was characterized by PCR

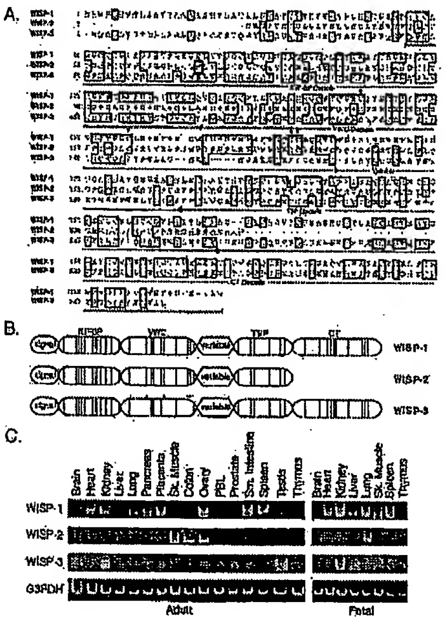


Fig. 3. (*A*) Encoded amino acid sequence alignment of human WISPs. The cysteine residues of WISP-1 and WISP-2 that are not present in WISP-3 are indicated with a dot. (*B*) Schematic representation of the WISP proteins showing the domain structure and cysteine residues (vertical lines). The four cysteine residues in the VWC domain that are absent in WISP-3 are indicated with a dot. (*C*) Expression of *WISP* mRNA in human tissues. PCR was performed on human multiple-tissue cDNA panels (CLONTECH) from the indicated adult and fetal tissues.

analysis on adult and fetal multiple tissue cDNA panels. *WISP-1* expression was seen in the adult heart, kidney, lung, pancreas, placenta, ovary, small intestine, and spleen (Fig. 3*C*). Little or no expression was detected in the brain, liver, skeletal muscle, colon, peripheral blood leukocytes, prostate, testis, or thymus. *WISP-2* had a more restricted tissue expression and was detected in adult skeletal muscle, colon, ovary, and fetal lung. Predominant expression of *WISP-3* was seen in adult kidney and testis and fetal kidney. Lower levels of *WISP-3* expression were detected in placenta, ovary, prostate, and small intestine.

***In Situ* Localization of *WISP-1* and *WISP-2*.** Expression of *WISP-1* and *WISP-2* was assessed by *in situ* hybridization in mammary tumors from Wnt-1 transgenic mice. Strong expression of *WISP-1* was observed in stromal fibroblasts lying within the fibrovascular tumor stroma (Fig. 4 *A–D*). However, low-level *WISP-1* expression also was observed focally within tumor cells (data not shown). No expression was observed in normal breast. Like *WISP-1*, *WISP-2* expression also was seen in the tumor stroma in breast tumors from Wnt-1 transgenic animals (Fig. 4 *E–H*). However, *WISP-2* expression in the stroma was in spindle-shaped cells adjacent to capillary vessels, whereas the predominant cell type expressing *WISP-1* was the stromal fibroblasts.

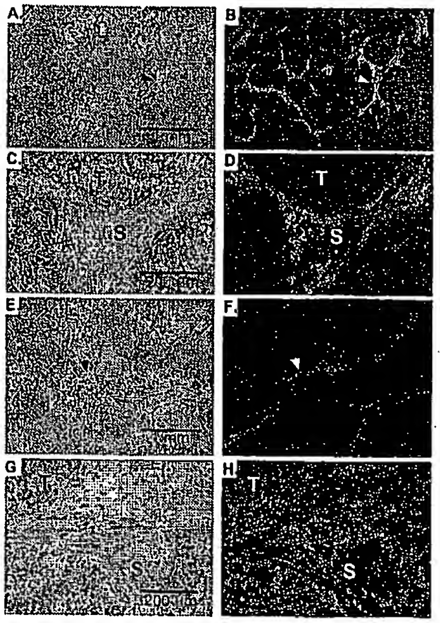


FIG. 4. (*A*, *C*, *E*, and *G*) Representative hematoxylin/eosin-stained images from breast tumors in Wnt-1 transgenic mice. The corresponding dark-field images showing *WISP-1* expression are shown in *B* and *D*. The tumor is a moderately well-differentiated adenocarcinoma showing evidence of adenoid cystic change. At low power (*A* and *B*), expression of *WISP-1* is seen in the delicate branching fibrovascular tumor stroma (arrowhead). At higher magnification, expression is seen in the stromal(s) fibroblasts (*C* and *D*), and tumor cells are negative. Focal expression of *WISP-1*, however, was observed in tumor cells in some areas. Images of *WISP-2* expression are shown in *E–H*. At low power (*E* and *F*), expression of WISP-2 is seen in cells lying within the fibrovascular tumor stroma. At higher magnification, these cells appeared to be adjacent to capillary vessels whereas tumor cells are negative (*G* and *H*).

Chromosome Localization of the *WISP* Genes. The chromosomal location of the human *WISP* genes was determined by radiation hybrid mapping panels. *WISP-1* is approximately 3.48 cR from the meiotic marker AFM259xc5 [logarithm of odds (lod) score 16.31] on chromosome 8q24.1 to 8q24.3, in the same region as the human locus of the *nov*H family member (27) and roughly 4 Mbs distal to c-*myc* (28). Preliminary fine mapping indicates that *WISP-1* is located near D8S1712 STS. *WISP-2* is linked to the marker SHGC-33922 (lod = 1,000) on chromosome 20q12–20q13.1. Human *WISP-3* mapped to chromosome 6q22–6q23 and is linked to the marker AFM211ze5 (lod = 1,000). *WISP-3* is approximately 18 Mbs proximal to CTGF and 23 Mbs proximal to the human cellular oncogene *MYB* (27, 29).

Amplification and Aberrant Expression of *WISPs* in Human Colon Tumors. Amplification of protooncogenes is seen in many human tumors and has etiological and prognostic significance. For example, in a variety of tumor types, c-*myc* amplification has been associated with malignant progression and poor prognosis (30). Because *WISP-1* resides in the same general chromosomal location (8q24) as c-*myc*, we asked whether it was a target of gene amplification, and, if so, whether this amplification was independent of the c-*myc* locus. Genomic DNA from human colon cancer cell lines was assessed by quantitative PCR and Southern blot analysis. (Fig. 5 *A* and *B*). Both methods detected similar degrees of *WISP-1* amplification. Most cell lines showed significant (2- to 4-fold) amplification, with the HT-29 and WiDr cell lines demonstrating an 8-fold increase. Significantly, the pattern of amplification observed did not correlate with that observed for c-*myc*, indicating that the c-*myc* gene is not part of the amplicon that involves the *WISP-1* locus.

We next examined whether the *WISP* genes were amplified in a panel of 25 primary human colon adenocarcinomas. The relative *WISP* gene copy number in each colon tumor DNA was compared with pooled normal DNA from 10 donors by quantitative PCR (Fig. 6). The copy number of *WISP-1* and *WISP-2* was significantly greater than one, approximately 2-fold for *WISP-1* in about 60% of the tumors and 2- to 4-fold for *WISP-2* in 92% of the tumors ($P < 0.001$ for each). The copy number for *WISP-3* was indistinguishable from one ($P = 0.166$). In addition, the copy number of *WISP-2* was significantly higher than that of *WISP-1* ($P < 0.001$).

The levels of *WISP* transcripts in RNA isolated from 19 adenocarcinomas and their matched normal mucosa were

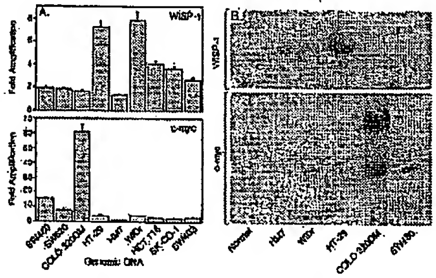


FIG. 5. Amplification of *WISP-1* genomic DNA in colon cancer cell lines. (*A*) Amplification in cell line DNA was determined by quantitative PCR. (*B*) Southern blots containing genomic DNA (10 μg) digested with *Eco*R1 (*WISP-1*) or *Xba*I (c-*myc*) were hybridized with a 100-bp human *WISP-1* probe (amino acids 186–219) or a human c-*myc* probe (located at bp 1901–2000). The *WISP* and *myc* genes are detected in normal human genomic DNA after a longer film exposure.

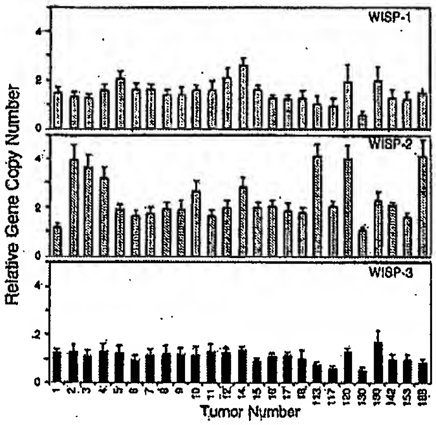


FIG. 6. Genomic amplification of *WISP* genes in human colon tumors. The relative gene copy number of the *WISP* genes in 25 adenocarcinomas was assayed by quantitative PCR, by comparing DNA from primary human tumors with pooled DNA from 10 healthy donors. The data are means ± SEM from one experiment done in triplicate. The experiment was repeated at least three times.

assessed by quantitative PCR (Fig. 7). The level of *WISP-1* RNA present in tumor tissue varied but was significantly increased (2- to >25-fold) in 84% (16/19) of the human colon tumors examined compared with normal adjacent mucosa. Four of 19 tumors showed greater than 10-fold overexpression. In contrast, in 79% (15/19) of the tumors examined, *WISP-2* RNA expression was significantly lower in the tumor than the mucosa. Similar to *WISP-1*, *WISP-3* RNA was overexpressed in 63% (12/19) of the colon tumors compared with the normal mucosa. The amount of overexpression of *WISP-3* ranged from 4- to >40-fold.

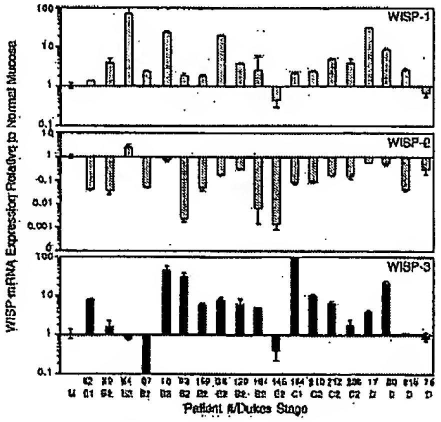


FIG. 7. *WISP* RNA expression in primary human colon tumors relative to expression in normal mucosa from the same patient. Expression of *WISP* mRNA in 19 adenocarcinomas was assayed by quantitative PCR. The Dukes stage of the tumor is listed under the sample number. The data are means ± SEM from one experiment done in triplicate. The experiment was repeated at least twice.

DISCUSSION

One approach to understanding the molecular basis of cancer is to identify differences in gene expression between cancer cells and normal cells. Strategies based on assumptions that steady-state mRNA levels will differ between normal and malignant cells have been used to clone differentially expressed genes (31). We have used a PCR-based selection strategy, SSH, to identify genes selectively expressed in C57MG mouse mammary epithelial cells transformed by Wnt-1.

Three of the genes isolated, *WISP-1*, *WISP-2*, and *WISP-3*, are members of the CCN family of growth factors, which includes CTGF, Cyr61, and *nov*, a family not previously linked to Wnt signaling.

Two independent experimental systems demonstrated that *WISP* induction was associated with the expression of Wnt-1. The first was C57MG cells infected with a Wnt-1 retroviral vector or C57MG cells expressing Wnt-1 under the control of a tetracyline-repressible promoter, and the second was in Wnt-1 transgenic mice, where breast tissue expresses Wnt-1, whereas normal breast tissue does not. No *WISP* RNA expression was detected in mammary tumors induced by polyoma virus middle T antigen (data not shown). These data suggest a link between Wnt-1 and *WISPs* in that in these two situations, *WISP* induction was correlated with Wnt-1 expression.

It is not clear whether the *WISPs* are directly or indirectly induced by the downstream components of the Wnt-1 signaling pathway (i.e., β-catenin-TCF-1/Lef1). The increased levels of *WISP* RNA were measured in Wnt-1-transformed cells, hours or days after Wnt-1 transformation. Thus, *WISP* expression could result from Wnt-1 signaling directly through β-catenin transcription factor regulation or alternatively through Wnt-1 signaling turning on a transcription factor, which in turn regulates *WISPs*.

The *WISPs* define an additional subfamily of the CCN family of growth factors. One striking difference observed in the protein sequence of WISP-2 is the absence of a CT domain, which is present in CTGF, Cyr61, *nov*, WISP-1, and WISP-3. This domain is thought to be involved in receptor binding and dimerization. Growth factors, such as TGF-β, platelet-derived growth factor, and nerve growth factor, which contain a cystine knot motif exist as dimers (32). It is tempting to speculate that WISP-1 and WISP-3 may exist as dimers, whereas WISP-2 exists as a monomer. If the CT domain is also important for receptor binding, WISP-2 may bind its receptor through a different region of the molecule than the other CCN family members. No specific receptors have been identified for CTGF or *nov*. A recent report has shown that integrin $\alpha_v\beta_3$ serves as an adhesion receptor for Cyr61 (33).

The strong expression of *WISP-1* and *WISP-2* in cells lying within the fibrovascular tumor stroma in breast tumors from Wnt-1 transgenic animals is consistent with previous observations that transcripts for the related CTGF gene are primarily expressed in the fibrous stroma of mammary tumors (34). Epithelial cells are thought to control the proliferation of connective tissue stroma in mammary tumors by a cascade of growth factor signals similar to that controlling connective tissue formation during wound repair. It has been proposed that mammary tumor cells or inflammatory cells at the tumor interstitial interface secrete TGF-β1, which is the stimulus for stromal proliferation (34). TGF-β1 is secreted by a large percentage of malignant breast tumors and may be one of the growth factors that stimulates the production of CTGF and WISPs in the stroma.

It was of interest that *WISP-1* and *WISP-2* expression was observed in the stromal cells that surrounded the tumor cells (epithelial cells) in the Wnt-1 transgenic mouse sections of breast tissue. This finding suggests that paracrine signaling could occur in which the stromal cells could supply WISP-1 and WISP-2 to regulate tumor cell growth on the WISP extracellular matrix. Stromal cell-derived factors in the extracellular matrix have been postulated to play a role in tumor cell migration and proliferation (35). The localization of *WISP-1* and *WISP-2* in the stromal cells of breast tumors supports this paracrine model.

An analysis of *WISP-1* gene amplification and expression in human colon tumors showed a correlation between DNA amplification and overexpression, whereas overexpression of *WISP-3* RNA was seen in the absence of DNA amplification. In contrast, *WISP-2* DNA was amplified in the colon tumors, but its mRNA expression was significantly reduced in the majority of tumors compared with the expression in normal colonic mucosa from the same patient. The gene for human *WISP-2* was localized to chromosome 20q12–20q13, at a region frequently amplified and associated with poor prognosis in node negative breast cancer and many colon cancers, suggesting the existence of one or more oncogenes at this locus (36–38). Because the center of the 20q13 amplicon has not yet been identified, it is possible that the apparent amplification observed for *WISP-2* may be caused by another gene in this amplicon.

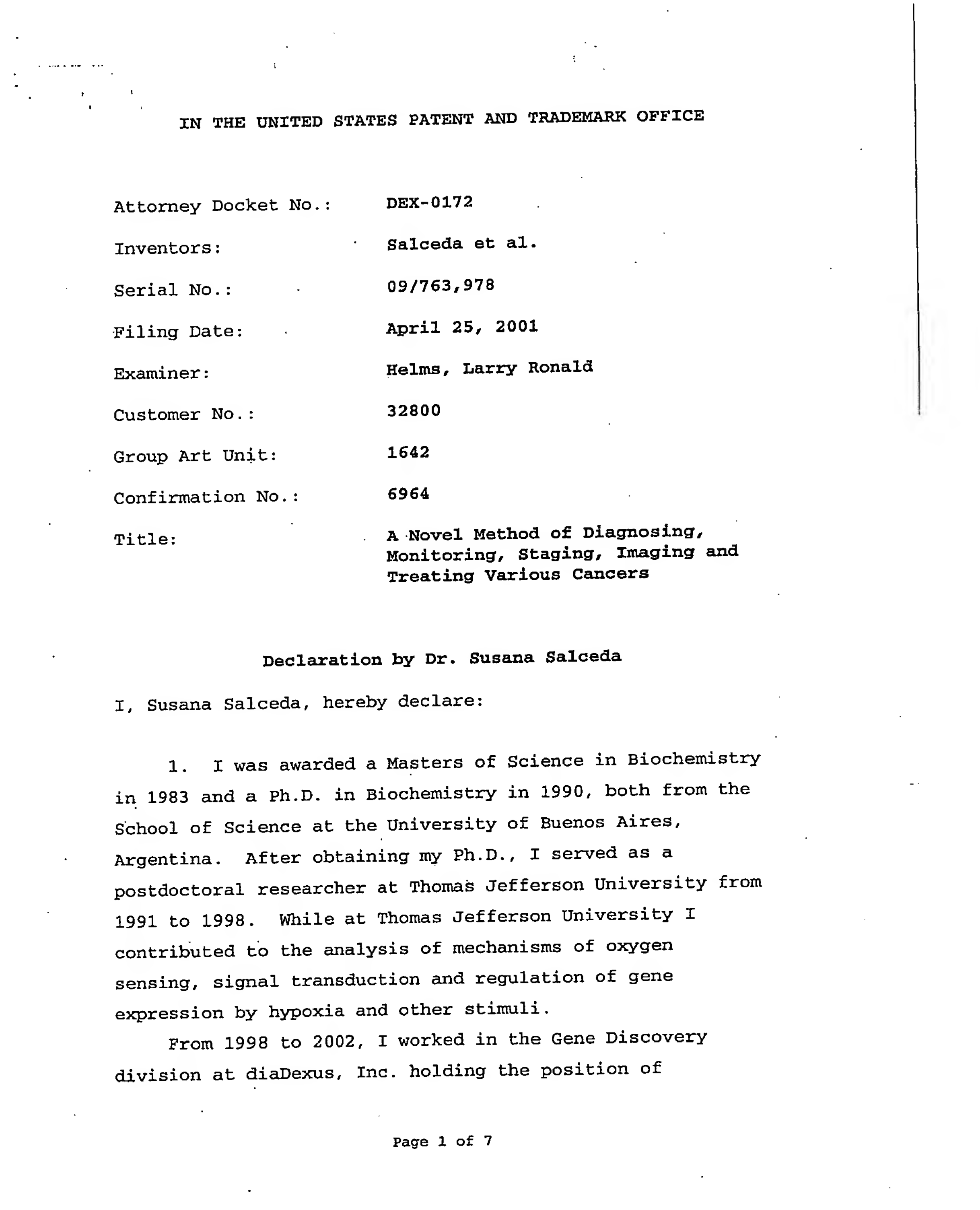
A recent manuscript on *rCop-1*, the rat orthologue of *WISP-2*, describes the loss of expression of this gene after cell transformation, suggesting it may be a negative regulator of growth in cell lines (16). Although the mechanism by which *WISP-2* RNA expression is down-regulated during malignant transformation is unknown, the reduced expression of *WISP-2* in colon tumors and cell lines suggests that it may function as a tumor suppressor. These results show that the *WISP* genes are aberrantly expressed in colon cancer and suggest that their altered expression may confer selective growth advantage to the tumor.

Members of the Wnt signaling pathway have been implicated in the pathogenesis of colon cancer, breast cancer, and melanoma, including the tumor suppressor gene adenomatous polyposis coli and β-catenin (39). Mutations in specific regions of either gene can cause the stabilization and accumulation of cytoplasmic β-catenin, which presumably contributes to human carcinogenesis through the activation of target genes such as the *WISP*s. Although the mechanism by which Wnt-1 transforms cells and induces tumorigenesis is unknown, the identification of *WISP*s as genes that may be regulated downstream of Wnt-1 in C57MG cells suggests they could be important mediators of Wnt-1 transformation. The amplification and altered expression patterns of the *WISP*s in human colon tumors may indicate an important role for these genes in tumor development.

We thank the DNA synthesis group for oligonucleotide synthesis, T. Baker for technical assistance, P. Dowd for radiation hybrid mapping, K. Willert and R. Nusse for the tet-repressible C57MG/Wnt-1 cells, V. Dixit for discussions, and D. Wood and A. Bruce for artwork.

1. Cadigan, K. M. & Nusse, R. (1997) *Genes Dev.* 11, 3286–3305.
2. Dale, T. C. (1998) *Biochem. J.* 329, 209–223.
3. Nusse, R. & Varmus, H. E. (1982) *Cell* 31, 99–109.
4. van Ooyen, A. & Nusse, R. (1984) *Cell* 39, 233–240.
5. Tsukamoto, A. S., Grosschedl, R., Guzman, R. C., Parslow, T. & Varmus, H. E. (1988) *Cell* 55, 619–625.
6. Brown, J. D. & Moon, R. T. (1998) *Curr. Opin. Cell. Biol.* 10, 182–187.
7. Molenaar, M., van de Wetering, M., Oosterwegel, M., Peterson-Maduro, J., Godsave, S., Korinek, V., Roose, J., Destree, O. & Clevers, H. (1996) *Cell* 86, 391–399.
8. Korinek, V., Barker, N., Willert, K., Molenaar, M., Roose, J., Wagenaar, G., Markman, M., Lamers, W., Destree, O. & Clevers, H. (1998) *Mol. Cell. Biol.* 18, 1248–1256.
9. Munemitsu, S., Albert, I., Souza, B., Rubinfeld, B. & Polakis, P. (1995) *Proc. Natl. Acad. Sci. USA* 92, 3046–3050.
10. He, T. C., Sparks, A. B., Rago, C., Hermeking, H., Zawel, L., da Costa, L. T., Morin, P. J., Vogelstein, B. & Kinzler, K. W. (1998) *Science* 281, 1509–1512.
11. Diatchenko, L., Lau, Y. F., Campbell, A. P., Chenchik, A., Moqadam, F., Huang, B., Lukyanov, S., Lukyanov, K., Gurskaya, N., Sverdlov, E. D. & Siebert, P. D. (1996) *Proc. Natl. Acad. Sci. USA* 93, 6025–6030.
12. Brown, A. M., Wildin, R. S., Prendergast, T. J. & Varmus, H. E. (1986) *Cell* 46, 1001–1009.
13. Wong, G. T., Gavin, B. J. & McMahon, A. P. (1994) *Mol. Cell. Biol.* 14, 6278–6286.
14. Shimizu, H., Julius, M. A., Giarre, M., Zheng, Z., Brown, A. M. & Kitajewski, J. (1997) *Cell Growth Differ.* 8, 1349–1358.
15. Hashimoto, Y., Shindo-Okada, N., Tani, M., Nagamachi, Y., Takeuchi, K., Shiroishi, T., Toma, H. & Yokota, J. (1998) *J. Exp. Med.* 187, 289–296.
16. Zhang, R., Averboukh, L., Zhu, W., Zhang, H., Jo, H., Dempsey, P. J., Coffey, R. J., Pardee, A. B. & Liang, P. (1998) *Mol. Cell. Biol.* 18, 6131–6141.
17. Grotendorst, G. R. (1997) *Cytokine Growth Factor Rev.* 8, 171–179.
18. Kireeva, M. L., Mo, F. E., Yang, G. P. & Lau, L. F. (1996) *Mol. Cell. Biol.* 16, 1326–1334.
19. Babic, A. M., Kireeva, M. L., Kolesnikova, T. V. & Lau, L. F. (1998) *Proc. Natl. Acad. Sci. USA* 95, 6355–6360.
20. Martinerie, C., Huff, V., Joubert, I., Badzioch, M., Saunders, G., Strong, L. & Perbal, B. (1994) *Oncogene* 9, 2729–2732.
21. Bork, P. (1993) *FEBS Lett.* 327, 125–130.
22. Kim, H. S., Nagalla, S. R., Oh, Y., Wilson, E., Roberts, C. T., Jr. & Rosenfeld, R. G. (1997) *Proc. Natl. Acad. Sci. USA* 94, 12981–12986.
23. Joliot, V., Martinerie, C., Dambrine, G., Plassiart, G., Brisac, M., Crochet, J. & Perbal, B. (1992) *Mol. Cell. Biol.* 12, 10–21.
24. Mancuso, D. J., Tuley, E. A., Westfield, L. A., Worrall, N. K., Shelton-Inloes, B. B., Sorace, J. M., Alevy, Y. G. & Sadler, J. E. (1989) *J. Biol. Chem.* 264, 19514–19527.
25. Holt, G. D., Pangburn, M. K. & Ginsburg, V. (1990) *J. Biol. Chem.* 265, 2852–2855.
26. Voorberg, J., Fontijn, R., Calafat, J., Janssen, H., van Mourik, J. A. & Pannekoek, H. (1991) *J. Cell. Biol.* 113, 195–205.
27. Martinerie, C., Viegas-Pequignot, E., Guenard, I., Dutrillaux, B., Nguyen, V. C., Bernheim, A. & Perbal, B. (1992) *Oncogene* 7, 2529–2534.
28. Takahashi, E., Hori, T., O'Connell, P., Leppert, M. & White, R. (1991) *Cytogenet. Cell. Genet.* 57, 109–111.
29. Meese, E., Meltzer, P. S., Witkowski, C. M. & Trent, J. M. (1989) *Genes Chromosomes Cancer* 1, 88–94.
30. Garte, S. J. (1993) *Crit. Rev. Oncog.* 4, 435–449.
31. Zhang, L., Zhou, W., Velculescu, V. E., Kern, S. E., Hruban, R. H., Hamilton, S. R., Vogelstein, B. & Kinzler, K. W. (1997) *Science* 276, 1268–1272.
32. Sun, P. D. & Davies, D. R. (1995) *Annu. Rev. Biophys. Biomol. Struct.* 24, 269–291.
33. Kireeva, M. L., Lam, S. C. T. & Lau, L. F. (1998) *J. Biol. Chem.* 273, 3090–3096.
34. Frazier, K. S. & Grotendorst, G. R. (1997) *Int. J. Biochem. Cell. Biol.* 29, 153–161.
35. Wernert, N. (1997) *Virchows Arch.* 430, 433–443.
36. Tanner, M. M., Tirkkonen, M., Kallioniemi, A., Collins, C., Stokke, T., Karhu, R., Kowbel, D., Shadravan, F., Hintz, M., Kuo, W. L., *et al.* (1994) *Cancer Res.* 54, 4257–4260.
37. Brinkmann, U., Gallo, M., Polymeropoulos, M. H. & Pastan, I. (1996) *Genome Res.* 6, 187–194.
38. Bischoff, J. R., Anderson, L., Zhu, Y., Mossie, K., Ng, L., Souza, B., Schryver, B., Flanagan, P., Clairvoyant, F., Ginther, C., *et al.* (1998) *EMBO J.* 17, 3052–3065.
39. Morin, P. J., Sparks, A. B., Korinek, V., Barker, N., Clevers, H., Vogelstein, B. & Kinzler, K. W. (1997) *Science* 275, 1787–1790.
40. Lu, L. H. & Gillett, N. (1994) *Cell Vision* 1, 169–176.

Appendix E



IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Attorney Docket No.:	**DEX-0172**
Inventors:	**Salceda et al.**
Serial No.:	**09/763,978**
Filing Date:	**April 25, 2001**
Examiner:	**Helms, Larry Ronald**
Customer No.:	**32800**
Group Art Unit:	**1642**
Confirmation No.:	**6964**
Title:	**A Novel Method of Diagnosing, Monitoring, Staging, Imaging and Treating Various Cancers**

Declaration by Dr. Susana Salceda

I, Susana Salceda, hereby declare:

1. I was awarded a Masters of Science in Biochemistry in 1983 and a Ph.D. in Biochemistry in 1990, both from the School of Science at the University of Buenos Aires, Argentina. After obtaining my Ph.D., I served as a postdoctoral researcher at Thomas Jefferson University from 1991 to 1998. While at Thomas Jefferson University I contributed to the analysis of mechanisms of oxygen sensing, signal transduction and regulation of gene expression by hypoxia and other stimuli.

From 1998 to 2002, I worked in the Gene Discovery division at diaDexus, Inc. holding the position of

Scientist. At diaDexus I contributed to research using genomics based analyses focusing on the discovery, identification and characterization of novel polynucleotides and encoded proteins differentially expressed in cancer. Identified polynucleotides and encoded proteins were used to develop novel diagnostic and therapeutic products for the improved detection, classification, prognosis and treatment of cancer.

Since 2002, I have been a Senior Scientist working in the Expression Product Development Department at Affymetrix, Inc., in Santa Clara, CA. At Affymetrix I contribute to the development of new assays and reagents to process DNA and RNA samples for microarray analysis.

2. As a scientist, a former diaDexus employee, and a named inventor, I am familiar with the teachings of the above-referenced patent application. I was responsible for the discovery of Ovr110 and the sequences encoding it.

3. I have reviewed and am familiar with the office action in the above-referenced patent application dated June 22, 2005 from the U.S. Patent Office.

4. I understand the Examiner has taken a position that the "invention is not supported by either a substantial asserted utility or a well established utility." I respectfully disagree.

5. At the time of the invention the usefulness of an isolated antibody or antibody fragment that binds specifically to a cancer marker such as the protein encoded by polynucleotide SEQ ID NO: 1 was well known.

6. Further, at the time of the invention we routinely obtained a protein sequence or open reading frame from information related to a polynucleotide sequence such as that provided for the polynucleotide sequence of SEQ ID NO: 1.

For example, as shown in Examples 1 and 2 of the above referenced patent application, the sequence and expression data of SEQ ID NO:1 is based on an mRNA molecule and therefore has a set 5′ to 3′ orientation. Thus, from this information, we know the protein is encoded in the forward (5′ to 3′) direction of SEQ ID NO: 1.

Furthermore, since expressed mRNA encode for proteins we know that the open reading frame in the forward direction of SEQ ID NO: 1 would be in a frame encoding for a Methionine near the 5′ end, encode many amino acids and terminate with a stop codon. Thus, any reading frame sequence of SEQ ID NO: 1 with lots of stop codons can be ruled out since we know to look for a long open reading frame sequence beginning with an M and ending with a stop codon in accordance with the information taught in the patent application about SEQ ID NO: 1.

By 1998 there were many tools available for use to determine either the protein sequence or the open reading frame (ORF) of a sequence such as SEQ ID NO: 1. Examples of such programs include the MAP[1] application, part of the GCG software suite from Accelrys Software Inc. (San Diego, CA), the Translate application, part of ExPASy (Expert Protein Analysis System) available online (at www.expasy.org/tools/dna.html) from the Swiss Institute of

[1] Devereux J, Haeberli P, Smithies O. (1984 NAR 11, 387-395)

Bioinformatics (Lausanne, Switzerland) and the ORF Finder (Open Reading Frame Finder) application available online (at www.ncbi.nlm.nih.gov/gorf/gorf.html)from the National Center for Biotechnology Information (NCBI) (Bethesda, MD).

As examples, attached are the results of the MAP, Translate and ORF Finder programs described above. The attached MAP program results (Figure 1) display SEQ ID NO: 1 as taught in the patent application in the forward direction, the reverse complement strand, and the protein translation of the three frames of the forward nucleotide strand followed by the protein translation of the three frames of the reverse compliment strand. For clarity, the open reading frame and protein encoded by SEQ ID NO: 1 have been underlined. As with many programs, the start codons encoding a Methionine (denoted by "**M**" or "**Met**") and stop codons not encoding an amino acid (denoted by "*****" or "**Stop**") are in bold. Also displayed in the MAP results, but not relevant to the open reading frame or encoded protein, are the nucleotide restriction sites for the endonuclease SAU3AI.

The attached Translate program results (Figure 2) display the protein translations of the three forward frames (5'3') followed by the protein translation of the three frames of the reverse compliment strand (3'5'). For clarity, the protein encoded by SEQ ID NO: 1 has been underlined.

The attached ORF Finder program results (Figure 3) displays a graphical representation of the ORFs greater than 100 nucleotides in length in each of the six frames of SEQ ID NO: 1. The longest open reading frame is listed first on the right as frame +2 from nucleotide 62-910 with a length of 849 nucleotides. This open reading frame is selected (highlighted) in the display and the ORF nucleotide sequence and encoded 282 amino acid protein sequence is displayed below.

Using the attached results from the MAP application, Translate application, ORF Finder application, or output from another simple translation program, the encoded protein and open reading frame are clear. Here MAP, Translate or ORF Finder show the protein encoded by SEQ ID NO: 1 is 282 amino acids long. Thus, using only the information taught in the specification as filed, the open reading frame for SEQ ID NO: 1 and the encoded protein can be routinely and unambiguously identified.

7. The Examiner also suggests that there was "no indication of what the protein [encoded by SEQ ID NO: 1] was." I respectfully disagree. As shown by the attached results from the MAP application, Translate application and ORF Finder application, the protein encoded by SEQ ID NO: 1 was readily obtainable with tools used routinely as of 1998.

8. Similarly, the process of expressing the protein encoded by a nucleotide such as SEQ ID NO: 1 and generating antibodies to the protein was well known as of 1998 and prior thereto.

9. I respectfully disagree with the Examiner's suggestion that this sequence and invention are "starting points for further research and investigation into potential practical uses." As shown herein, the nucleotide sequence of SEQ ID NO: 1 and the characteristics disclosed in the patent application about SEQ ID NO: 1 were adequate

to routinely and unambiguously obtain the protein sequence and then generate antibodies or antibody fragments thereto.

10. I also respectfully disagree with the Examiner's suggestions that "one would not have known a utility for such a protein" and that the "specification does not teach a utility for use of the antibody." The patent application teaches that "the mRNA overexpression in most of the matching samples tested are indicative of Ovr110... being a diagnostic marker for gynecologic cancers." Further, uses for the protein expressed by the CSG encoded by SEQ ID NO: 1 are explicitly described in the specification. Since the mRNA of SEQ ID NO: 1 is overexpressed in gynecologic cancers samples, and encodes a protein, the value of antibodies to this protein to detect overexpressed protein in gynecologic cancers would also be understood.

Further, the specification explicitly teaches that antibodies against Cancer Specific Genes (CSG) such as SEQ ID NO: 1 "can be used to detect or image localization of CSG in a patient for the purpose of detecting or diagnosing selected cancers."

The specification also explicitly teaches that antibodies against Cancer Specific Genes (CSG) such as SEQ ID NO: 1 "can be injected into a patient suspected of having a selected cancer for diagnostic and/or therapeutic purposes."

Furthermore, contrary to the Examiner's suggestion, the specification provides detailed teachings as to how one of skill in the art could use these antibodies in an ELISA assay or a competition assay to detect cancer, thus providing guidance regarding use of the invention "in a manner that constitutes a substantial utility."

11. I also respectfully disagree with the Examiner's suggestion that "applicants were not in possession of any protein encoded by SEQ ID NO: 1." As I showed herein, using standard tools available at the time of the invention, one of skill in the art could readily determine the protein encoded by SEQ ID NO: 1. All the necessary information to do so is provided by the polynucleotide sequence and the characteristics of this sequence taught in the patent application.

I hereby declare that all statements herein of my own knowledge are true and that all statements made on information or belief are believed to be true; and further that these statements were made with the knowledge that willful statements and the like so made are punishable by fine or by imprisonment, or both, under §1001 of Title 18 of the United States code, and that such willful statements may jeopardize the validity of the application, any patent issuing there upon, or any patent to which this verified statement is directed.

Susana Salceda,. Ph.D.

10/18/05

Date

FIGURE 1

1 / 9

```
(Linear) MAP of: dex0043_1.seq  check: 5695  from: 1  to: 2587

DEX0043_1

 With 1 enzymes: SAU3AI

Forward frame translations:

            ggaaggcagcgggcagctccactcagccagtacccagatacgctgggaaccttccccagc
          1 ---------+---------+---------+---------+---------+---------+ 60
            ccttccgtcgcccgtcgaggtgagtcggtcatgggtctatgcgacccttggaaggggtcg

a           G  R  Q  R  A  A  P  L  S  Q  Y  P  D  T  L  G  T  F  P  S   -
b            E  G  S  G  Q  L  H  S  A  S  T  Q  I  R  W  E  P  S  P  A  -
c             K  A  A  G  S  S  T  Q  P  V  P  R  Y  A  G  N  L  P  Q  P -

                        Sau3AI
                              |
            catggcttccctggggcagatcctcttctggagcataattagcatcatcattattctggc
         61 ---------+---------+---------+---------+---------+---------+ 120
            gtaccgaagggaccccgtctaggagaagacctcgtattaatcgtagtagtaataagaccg

a           H  G  F  P  G  A  D  P  L  L  E  H  N  *  H  H  H  Y  S  G   -
b            M  A  S  L  G  Q  I  L  F  W  S  I  I  S  I  I  I  I  L  A  -
c             W  L  P  W  G  R  S  S  S  G  A  *  L  A  S  S  L  F  W  L -

            tggagcaattgcactcatcattggctttggtatttcagggagacactccatcacagtcac
        121 ---------+---------+---------+---------+---------+---------+ 180
            acctcgttaacgtgagtagtaaccgaaaccataaagtccctctgtgaggtagtgtcagtg

a           W  S  N  C  T  H  H  W  L  W  Y  F  R  E  T  L  H  H  S  H   -
b            G  A  I  A  L  I  I  G  F  G  I  S  G  R  H  S  I  T  V  T  -
c             E  Q  L  H  S  S  L  A  L  V  F  Q  G  D  T  P  S  Q  S  L -

            tactgtcgcctcagctgggaacattggggaggatggaatcctgagctgcacttttgaacc
        181 ---------+---------+---------+---------+---------+---------+ 240
            atgacagcggagtcgacccttgtaacccctcctaccttaggactcgacgtgaaaacttgg

a           Y  C  R  L  S  W  E  H  W  G  G  W  N  P  E  L  H  F  *  T   -
b            T  V  A  S  A  G  N  I  G  E  D  G  I  L  S  C  T  F  E  P  -
c             L  S  P  Q  L  G  T  L  G  R  M  E  S  *  A  A  L  L  N  L -

            tgacatcaaactttctgatatcgtgatacaatggctgaaggaaggtgttttaggcttggt
        241 ---------+---------+---------+---------+---------+---------+ 300
            actgtagtttgaaagactatagcactatgttaccgacttccttccacaaaatccgaacca

a           *  H  Q  T  F  *  Y  R  D  T  M  A  E  G  R  C  F  R  L  G   -
b            D  I  K  L  S  D  I  V  I  Q  W  L  K  E  G  V  L  G  L  V  -
c             T  S  N  F  L  I  S  *  Y  N  G  *  R  K  V  F  *  A  W  S -

            ccatgagttcaaagaaggcaaagatgagctgtcggagcaggatgaaatgttcagaggccg
        301 ---------+---------+---------+---------+---------+---------+ 360
            ggtactcaagtttcttccgtttctactcgacagcctcgtcctactttacaagtctccggc

a           P  *  V  Q  R  R  Q  R  *  A  V  G  A  G  *  N  V  Q  R  P   -
b            H  E  F  K  E  G  K  D  E  L  S  E  Q  D  E  M  F  R  G  R  -
c             M  S  S  K  K  A  K  M  S  C  R  S  R  M  K  C  S  E  A  G -
```

FIGURE 1

```
                     Sau3AI
                        |
      gacagcagtgtttgctgatcaagtgatagttggcaatgcctctttgcggctgaaaaacgt
  361 ---------+---------+---------+---------+---------+---------+ 420
      ctgtcgtcacaaacgactagttcactatcaaccgttacggagaaacgccgactttttgca

a     D  S  S  V  C  *  S  S  D  S  W  Q  C  L  F  A  A  E  K  R   -
b      T  A  V  F  A  D  Q  V  I  V  G  N  A  S  L  R  L  K  N  V  -
c       Q  Q  C  L  L  I  K  *  *  L  A  M  P  L  C  G  *  K  T  C -

      gcaactcacagatgctggcacctacaaatgttatatcatcacttctaaaggcaaggggaa
  421 ---------+---------+---------+---------+---------+---------+ 480
      cgttgagtgtctacgaccgtggatgtttacaatatagtagtgaagatttccgttcccctt

a     A  T  H  R  C  W  H  L  Q  M  L  Y  H  H  F  *  R  Q  G  E   -
b      Q  L  T  D  A  G  T  Y  K  C  Y  I  I  T  S  K  G  K  G  N  -
c       N  S  Q  M  L  A  P  T  N  V  I  S  S  L  L  K  A  R  G  M -

      tgctaaccttgagtataaaactggagccttcagcatgccggaagtgaatgtggactataa
  481 ---------+---------+---------+---------+---------+---------+ 540
      acgattggaactcatattttgacctcggaagtcgtacggccttcacttacacctgatatt

a     C  *  P  *  V  *  N  W  S  L  Q  H  A  G  S  E  C  G  L  *   -
b      A  N  L  E  Y  K  T  G  A  F  S  M  P  E  V  N  V  D  Y  N  -
c       L  T  L  S  I  K  L  E  P  S  A  C  R  K  *  M  W  T  I  M -

      tgccagctcagagaccttgcggtgtgaggctccccgatggttcccccagcccacagtggt
  541 ---------+---------+---------+---------+---------+---------+ 600
      acggtcgagtctctggaacgccacactccgaggggctaccaagggggtcgggtgtcacca

a     C  Q  L  R  D  L  A  V  *  G  S  P  M  V  P  P  A  H  S  G   -
b      A  S  S  E  T  L  R  C  E  A  P  R  W  F  P  Q  P  T  V  V  -
c       P  A  Q  R  P  C  G  V  R  L  P  D  G  S  P  S  P  Q  W  S -

      ctgggcatcccaagttgaccagggagccaacttctcggaagtctccaataccagctttga
  601 ---------+---------+---------+---------+---------+---------+ 660
      gacccgtagggttcaactggtccctcggttgaagagccttcagaggttatggtcgaaact

a     L  G  I  P  S  *  P  G  S  Q  L  L  G  S  L  Q  Y  Q  L  *   -
b      W  A  S  Q  V  D  Q  G  A  N  F  S  E  V  S  N  T  S  F  E  -
c       G  H  P  K  L  T  R  E  P  T  S  R  K  S  P  I  P  A  L  S -

                                                          Sau3AI
                                                             |
      gctgaactctgagaatgtgaccatgaaggttgtgtctgtgctctacaatgttacgatcaa
  661 ---------+---------+---------+---------+---------+---------+ 720
      cgacttgagactcttacactggtacttccaacacagacacgagatgttacaatgctagtt

a     A  E  L  *  E  C  D  H  E  G  C  V  C  A  L  Q  C  Y  D  Q   -
b      L  N  S  E  N  V  T  M  K  V  V  S  V  L  Y  N  V  T  I  N  -
c       *  T  L  R  M  *  P  *  R  L  C  L  C  S  T  M  L  R  S  T -
```

FIGURE 1

3 / 9

```
      caacacatactcctgtatgattgaaaatgacattgccaaagcaacaggggatatcaaagt
 721  ---------+---------+---------+---------+---------+---------+  780
      gttgtgtatgaggacatactaacttttactgtaacggtttcgttgtcccctatagtttca

a     Q  H  I  L  L  Y  D  *  K  *  H  C  Q  S  N  R  G  Y  Q  S   -
b      N  T  Y  S  C  M  I  E  N  D  I  A  K  A  T  G  D  I  K  V  -
c       T  H  T  P  V  *  L  K  M  T  L  P  K  Q  Q  G  I  S  K  * -

          Sau3AI
              |
      gacagaatcggagatcaaaaggcggagtcacctacagctgctaaactcaaaggcttctct
 781  ---------+---------+---------+---------+---------+---------+  840
      ctgtcttagcctctagttttccgcctcagtggatgtcgacgatttgagtttccgaagaga

a     D  R  I  G  D  Q  K  A  E  S  P  T  A  A  K  L  K  G  F  S   -
b      T  E  S  E  I  K  R  R  S  H  L  Q  L  L  N  S  K  A  S  L  -
c       Q  N  R  R  S  K  G  G  V  T  Y  S  C  *  T  Q  R  L  L  C -

      gtgtgtctcttctttctttgccatcagctgggcacttctgcctctcagcccttacctgat
 841  ---------+---------+---------+---------+---------+---------+  900
      cacacagagaagaaagaaacggtagtcgacccgtgaagacggagagtcgggaatggacta

a     V  C  L  F  F  L  C  H  Q  L  G  T  S  A  S  Q  P  L  P  D   -
b      C  V  S  S  F  F  A  I  S  W  A  L  L  P  L  S  P  Y  L  M  -
c       V  S  L  L  S  L  P  S  A  G  H  F  C  L  S  A  L  T  *  C -

                                                        Sau3AI
                                                          |
      gctaaaataatgtgccttggccacaaaaaagcatgcaaagtcattgttacaacagggatc
 901  ---------+---------+---------+---------+---------+---------+  960
      cgattttattacacggaaccggtgtttttcgtacgtttcagtaacaatgttgtccctag

a     A  K  I  M  C  L  G  H  K  K  A  C  K  V  I  V  T  T  G  I   -
b      L  K  *  C  A  L  A  T  K  K  H  A  K  S  L  L  Q  Q  G  S  -
c       *  N  N  V  P  W  P  Q  K  S  M  Q  S  H  C  Y  N  R  D  L -

      tacagaactatttcaccaccagatatgacctagttttatatttctgggaggaaatgaatt
 961  ---------+---------+---------+---------+---------+---------+  1020
      atgtcttgataaagtggtggtctatactggatcaaaatataaagaccctcctttacttaa

a     Y  R  T  I  S  P  P  D  M  T  *  F  Y  I  S  G  R  K  *  I   -
b      T  E  L  F  H  H  Q  I  *  P  S  F  I  F  L  G  G  N  E  F  -
c       Q  N  Y  F  T  T  R  Y  D  L  V  L  Y  F  W  E  E  M  N  S -

      catatctagaagtctggagtgagcaaacaagagcaagaaacaaaaagaagccaaaagcag
1021  ---------+---------+---------+---------+---------+---------+  1080
      gtatagatcttcagacctcactcgtttgttctcgttctttgtttttcttcggttttcgtc

a     H  I  *  K  S  G  V  S  K  Q  E  Q  E  T  K  R  S  Q  K  Q   -
b      I  S  R  S  L  E  *  A  N  K  S  K  K  Q  K  E  A  K  S  R  -
c       Y  L  E  V  W  S  E  Q  T  R  A  R  N  K  K  K  P  K  A  E -
```

FIGURE 1

4 / 9

```
               aaggctccaatatgaacaagataaatctatcttcaaagacatattagaagttgggaaaat
          1081 ---------+---------+---------+---------+---------+---------+ 1140
               ttccgaggttatacttgttctatttagatagaagtttctgtataatcttcaacccttta

a              K  A  P  I  *  T  R  *  I  Y  L  Q  R  H  I  R  S  W  E  N   -
b               R  L  Q  Y  E  Q  D  K  S  I  F  K  D  I  L  E  V  G  K  I  -
c                G  S  N  M  N  K  I  N  L  S  S  K  T  Y  *  K  L  G  K  * -
```

FIGURE 1

5 / 9

```
Reverse frame translations:

        aattcatgtgaactagacaagtgtgttaagagtgataagtaaaatgcacgtggagacaag
   1141 ---------+---------+---------+---------+---------+---------+ 1200
        ttaagtacacttgatctgttcacacaattctcactattcattttacgtgcacctctgttc

a       N  S  C  E  L  D  K  C  V  K  S  D  K  *  N  A  R  G  D  K   -
b        I  H  V  N  *  T  S  V  L  R  V  I  S  K  M  H  V  E  T  S  -
c         F  M  *  T  R  Q  V  C  *  E  *  *  V  K  C  T  W  R  Q  V -

           Sau3AI
                |
        tgcatccccagatctcagggacctcccccctgcctgtcacctggggagtgagaggacagga
   1201 ---------+---------+---------+---------+---------+---------+ 1260
        acgtaggggtctagagtccctggaggggggacggacagtggacccctcactctcctgtcct

a       C  I  P  R  S  Q  G  P  P  P  A  C  H  L  G  S  E  R  T  G   -
b        A  S  P  D  L  R  D  L  P  L  P  V  T  W  G  V  R  G  Q  D  -
c         H  P  Q  I  S  G  T  S  P  C  L  S  P  G  E  *  E  D  R  I -

        tagtgcatgttctttgtctctgaattttagttatatgtgctgtaatgttgctctgagga
   1261 ---------+---------+---------+---------+---------+---------+ 1320
        atcacgtacaagaaacagagacttaaaatcaatatacacgacattacaacgagactcct

a       *  C  M  F  F  V  S  E  F  L  V  I  C  A  V  M  L  L  *  G   -
b        S  A  C  S  L  S  L  N  F  *  L  Y  V  L  *  C  C  S  E  E  -
c         V  H  V  L  C  L  *  I  F  S  Y  M  C  C  N  V  A  L  R  K -

        agcccctggaaagtctatcccaacatatccacatcttatattccacaaattaagctgtag
   1321 ---------+---------+---------+---------+---------+---------+ 1380
        tcggggacctttcagatagggttgtataggtgtagaatataaggtgtttaattcgacatc

a       S  P  W  K  V  Y  P  N  I  S  T  S  Y  I  P  Q  I  K  L  *   -
b        A  P  G  K  S  I  P  T  Y  P  H  L  I  F  H  K  L  S  C  S  -
c         P  L  E  S  L  S  Q  H  I  H  I  L  Y  S  T  N  *  A  V  V -

        tatgtaccctaagacgctgctaattgactgccacttcgcaactcaggggcggctgcattt
   1381 ---------+---------+---------+---------+---------+---------+ 1440
        atacatgggattctgcgacgattaactgacggtgaagcgttgagtccccgccgacgtaaa

a       Y  V  P  *  D  A  A  N  *  L  P  L  R  N  S  G  A  A  A  F   -
b        M  Y  P  K  T  L  L  I  D  C  H  F  A  T  Q  G  R  L  H  F  -
c         C  T  L  R  R  C  *  L  T  A  T  S  Q  L  R  G  G  C  I  L -

        tagtaatgggtcaaatgattcacttttttatgatgcttccaaaggtgccttggcttctctt
   1441 ---------+---------+---------+---------+---------+---------+ 1500
        atcattacccagtttactaagtgaaaaatactacgaaggtttccacggaaccgaagagaa

a       *  *  W  V  K  *  F  T  F  Y  D  A  S  K  G  A  L  A  S  L   -
b        S  N  G  S  N  D  S  L  F  M  M  L  P  K  V  P  W  L  L  F  -
c         V  M  G  Q  M  I  H  F  L  *  C  F  Q  R  C  L  G  F  S  S -
```

FIGURE 1

6 / 9

```
                                   Sau3AI
                                     |
     cccaactgacaaatgccaaagttgagaaaatgatcataattttagcataaacagagcag
1501 ---------+---------+---------+---------+---------+---------+ 1560
     gggttgactgtttacggtttcaactcttttactagtattaaaatcgtatttgtctcgtc

a    P  N  *  Q  M  P  K  L  R  K  M  I  I  I  L  A  *  T  E  Q   -
b     P  T  D  K  C  Q  S  *  E  K  *  S  *  F  *  H  K  Q  S  S  -
c      Q  L  T  N  A  K  V  E  K  N  D  H  N  F  S  I  N  R  A  V -

     tcggcgacaccgattttataaataaactgagcaccttcttttaaacaaacaaatgcggg
1561 ---------+---------+---------+---------+---------+---------+ 1620
     agccgctgtggctaaaatatttatttgactcgtggaagaaaatttgtttgtttacgccc

a    S  A  T  P  I  L  *  I  N  *  A  P  S  F  *  T  N  K  C  G   -
b     R  R  H  R  F  Y  K  *  T  E  H  L  L  F  K  Q  T  N  A  G  -
c      G  D  T  D  F  I  N  K  L  S  T  F  F  L  N  K  Q  M  R  V -

     tttatttctcagatgatgttcatccgtgaatggtccagggaaggacctttcaccttgact
1621 ---------+---------+---------+---------+---------+---------+ 1680
     aaataaagagtctactacaagtaggcacttaccaggtcccttcctggaaagtggaactga

a    F  I  S  Q  M  M  F  I  R  E  W  S  R  E  G  P  F  T  L  T   -
b     L  F  L  R  *  C  S  S  V  N  G  P  G  K  D  L  S  P  *  L  -
c      Y  F  S  D  D  V  H  P  *  M  V  Q  G  R  T  F  H  L  D  Y -

     atatggcattatgtcatcacaagctctgaggcttctcctttccatcctgcgtggacagct
1681 ---------+---------+---------+---------+---------+---------+ 1740
     tataccgtaatacagtagtgttcgagactccgaagaggaaaggtaggacgcacctgtcga

a    I  W  H  Y  V  I  T  S  S  E  A  S  P  F  H  P  A  W  T  A   -
b     Y  G  I  M  S  S  Q  A  L  R  L  L  L  S  I  L  R  G  Q  L  -
c      M  A  L  C  H  H  K  L  *  G  F  S  F  P  S  C  V  D  S  * -

     aagacctcagttttcaatagcatctagagcagtgggactcagctggggtgatttcgcccc
1741 ---------+---------+---------+---------+---------+---------+ 1800
     ttctggagtcaaaagttatcgtagatctcgtcaccctgagtcgaccccactaaagcgggg

a    K  T  S  V  F  N  S  I  *  S  S  G  T  Q  L  G  *  F  R  P   -
b     R  P  Q  F  S  I  A  S  R  A  V  G  L  S  W  G  D  F  A  P  -
c      D  L  S  F  Q  *  H  L  E  Q  W  D  S  A  G  V  I  S  P  P -

     ccatctccgggggaatgtctgaagacaattttggttacctcaatgagggagtggaggagg
1801 ---------+---------+---------+---------+---------+---------+ 1860
     ggtagaggcccccttacagacttctgttaaaaccaatggagttactccctcacctcctcc

a    P  S  P  G  E  C  L  K  T  I  L  V  T  S  M  R  E  W  R  R   -
b     H  L  R  G  N  V  *  R  Q  F  W  L  P  Q  *  G  S  G  G  G  -
c      I  S  G  G  M  S  E  D  N  F  G  Y  L  N  E  G  V  E  E  D -
```

FIGURE 1

7 / 9

```
          atacagtgctactaccaactagtggataaaggccagggatgctgctcaacctcctaccat
     1861 ---------+---------+---------+---------+---------+---------+ 1920
          tatgtcacgatgatggttgatcacctatttccggtccctacgacgagttggaggatggta

a         I  Q  C  Y  Y  Q  L  V  D  K  G  Q  G  C  C  S  T  S  Y  H   -
b          Y  S  A  T  T  N  *  W  I  K  A  R  D  A  A  Q  P  P  T  M  -
c           T  V  L  L  P  T  S  G  *  R  P  G  M  L  L  N  L  L  P  C -

          gtacaggacgtctccccattacaactacccaatccgaagtgtcaactgtgtcaggactaa
     1921 ---------+---------+---------+---------+---------+---------+ 1980
          catgtcctgcagaggggtaatgttgatgggttaggcttcacagttgacacagtcctgatt

a         V  Q  D  V  S  P  L  Q  L  P  N  P  K  C  Q  L  C  Q  D  *   -
b          Y  R  T  S  P  H  Y  N  Y  P  I  R  S  V  N  C  V  R  T  K  -
c           T  G  R  L  P  I  T  T  T  Q  S  E  V  S  T  V  S  G  L  R -

          gaaaccctggttttgagtagaaaagggcctggaaagaggggagccaacaaatctgtctgc
     1981 ---------+---------+---------+---------+---------+---------+ 2040
          ctttgggaccaaaactcatcttttcccggacctttctcccctcggttgtttagacagacg

a         E  T  L  V  L  S  R  K  G  P  G  K  R  G  A  N  K  S  V  C   -
b          K  P  W  F  *  V  E  K  G  L  E  R  G  E  P  T  N  L  S  A  -
c           N  P  G  F  E  *  K  R  A  W  K  E  G  S  Q  Q  I  C  L  L -

          ttctcacattagtcattggcaaataagcattctgtctctttggctgctgcctcagcacag
     2041 ---------+---------+---------+---------+---------+---------+ 2100
          aagagtgtaatcagtaaccgtttattcgtaagacagagaaaccgacgacggagtcgtgtc

a         F  S  H  *  S  L  A  N  K  H  S  V  S  L  A  A  A  S  A  Q   -
b          S  H  I  S  H  W  Q  I  S  I  L  S  L  W  L  L  P  Q  H  R  -
c           L  T  L  V  I  G  K  *  A  F  C  L  F  G  C  C  L  S  T  E -

          agagccagaactctatcgggcaccaggataacatctctcagtgaacagagttgacaaggc
     2101 ---------+---------+---------+---------+---------+---------+ 2160
          tctcggtcttgagatagcccgtggtcctattgtagagagtcacttgtctcaactgttccg

a         R  A  R  T  L  S  G  T  R  I  T  S  L  S  E  Q  S  *  Q  G   -
b          E  P  E  L  Y  R  A  P  G  *  H  L  S  V  N  R  V  D  K  A  -
c           S  Q  N  S  I  G  H  Q  D  N  I  S  Q  *  T  E  L  T  R  P -

          ctatgggaaatgcctgatgggattatcttcagcttgttgagcttctaagtttctttccct
     2161 ---------+---------+---------+---------+---------+---------+ 2220
          gataccctttacggactaccctaatagaagtcgaacaactcgaagattcaaagaaaggga

a         L  W  E  M  P  D  G  I  I  F  S  L  L  S  F  *  V  S  F  P   -
b          Y  G  K  C  L  M  G  L  S  S  A  C  *  A  S  K  F  L  S  L  -
c           M  G  N  A  *  W  D  Y  L  Q  L  V  E  L  L  S  F  F  P  F -

          tcattctaccctgcaagccaagttctgtaagagaaatgcctgagttctagctcaggtttt
     2221 ---------+---------+---------+---------+---------+---------+ 2280
          agtaagatgggacgttcggttcaagacattctctttacggactcaagatcgagtccaaaa

a         S  F  Y  P  A  S  Q  V  L  *  E  K  C  L  S  S  S  S  G  F   -
b          H  S  T  L  Q  A  K  F  C  K  R  N  A  *  V  L  A  Q  V  F  -
c           I  L  P  C  K  P  S  S  V  R  E  M  P  E  F  *  L  R  F  S -
```

FIGURE 1

8 / 9

```
                   Sau3AI
                        |
          cttactctgaatttagatctccagacccttcctggccacaattcaaattaaggcaacaaa
     2281 ---------+---------+---------+---------+---------+---------+ 2340
          gaatgagacttaaatctagaggtctgggaaggaccggtgttaagtttaattccgttgttt

a         L  T  L  N  L  D  L  Q  T  L  P  G  H  N  S  N  *  G  N  K   -
b          L  L  *  I  *  I  S  R  P  F  L  A  T  I  Q  I  K  A  T  N  -
c           Y  S  E  F  R  S  P  D  P  S  W  P  Q  F  K  L  R  Q  Q  T -

          catataccttccatgaagcacacacagacttttgaaagcaaggacaatgactgcttgaat
     2341 ---------+---------+---------+---------+---------+---------+ 2400
          gtatatggaaggtacttcgtgtgtgtctgaaaactttcgttcctgttactgacgaactta

a         H  I  P  S  M  K  H  T  Q  T  F  E  S  K  D  N  D  C  L  N   -
b          I  Y  L  P  *  S  T  H  R  L  L  K  A  R  T  M  T  A  *  I  -
c           Y  T  F  H  E  A  H  T  D  F  *  K  Q  G  Q  *  L  L  E  L -

          tgaggccttgaggaatgaagctttgaaggaaaagaatactttgtttccagccccccttccc
     2401 ---------+---------+---------+---------+---------+---------+ 2460
          actccggaactccttacttcgaaacttccttttcttatgaaacaaaggtcgggggaaggg

a         *  G  L  E  E  *  S  F  E  G  K  E  Y  F  V  S  S  P  L  P   -
b          E  A  L  R  N  E  A  L  K  E  K  N  T  L  F  P  A  P  F  P  -
c           R  P  *  G  M  K  L  *  R  K  R  I  L  C  F  Q  P  P  S  H -

          acactcttcatgtgttaaccactgccttcctggaccttggagccacggtgactgtattac
     2461 ---------+---------+---------+---------+---------+---------+ 2520
          tgtgagaagtacacaattggtgacggaaggacctggaacctcggtgccactgacataatg

a         T  L  F  M  C  *  P  L  P  S  W  T  L  E  P  R  *  L  Y  Y   -
b          H  S  S  C  V  N  H  C  L  P  G  P  W  S  H  G  D  C  I  T  -
c           T  L  H  V  L  T  T  A  F  L  D  L  G  A  T  V  T  V  L  H -

                                    Sau3AI
                                          |
          atgttgttatagaaaactgattttagagttctgatcgttcaagagaatgattaaatatac
     2521 ---------+---------+---------+---------+---------+---------+ 2580
          tacaacaatatcttttgactaaaatctcaagactagcaagttctcttactaatttatatg

a         M  L  L  *  K  T  D  F  R  V  L  I  V  Q  E  N  D  *  I  Y   -
b          C  C  Y  R  K  L  I  L  E  F  *  S  F  K  R  M  I  K  Y  T  -
c           V  V  I  E  N  *  F  *  S  S  D  R  S  R  E  *  L  N  I  H -

          atttcct
     2581 ------- 2587
          taaagga

a         I  S   -
b          F  P  -
c           F    -
```

FIGURE 1

9 / 9

Enzymes that do cut:

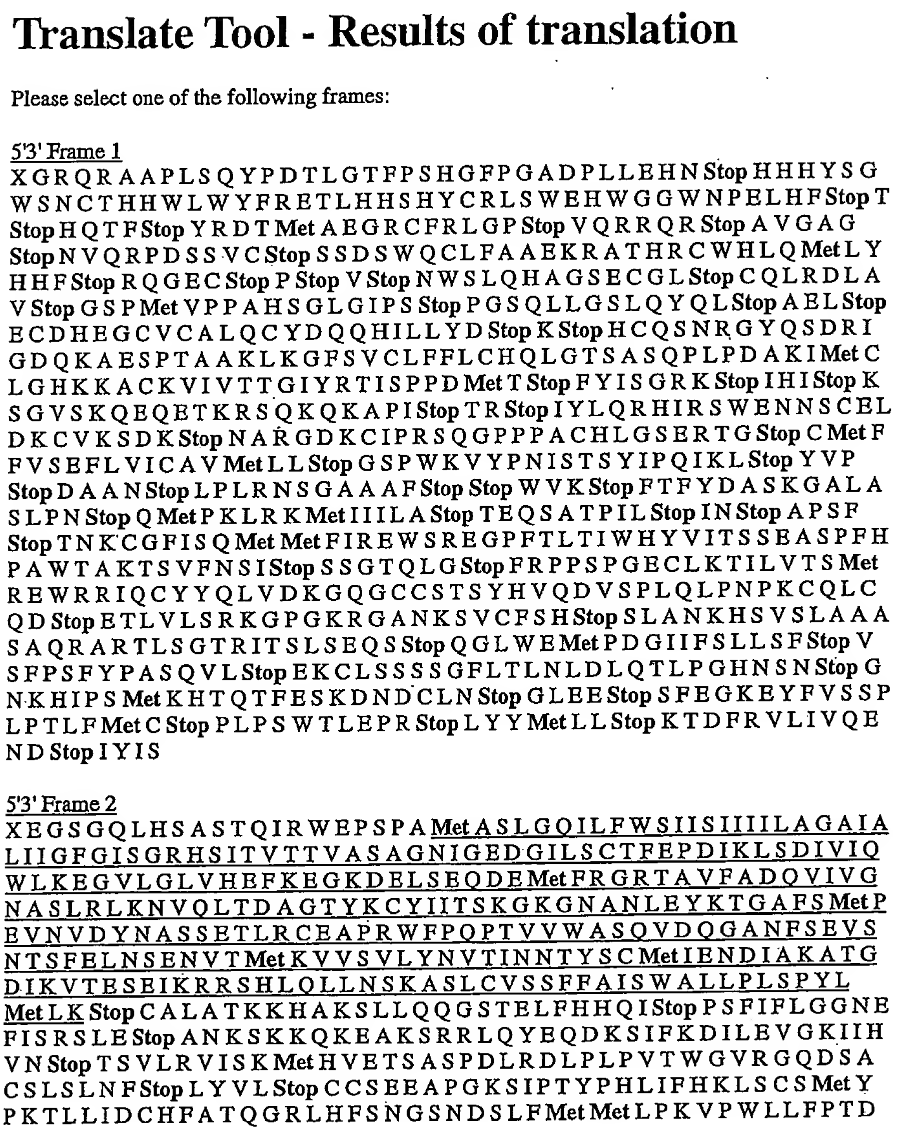
Sau3AI

Enzymes that do not cut:

NONE

FIGURE 2

1 / 4



Translate Tool - Results of translation

Please select one of the following frames:

5'3' Frame 1

X G R Q R A A P L S Q Y P D T L G T F P S H G F P G A D P L L E H N Stop H H H Y S G
W S N C T H H W L W Y F R E T L H H S H Y C R L S W E H W G G W N P E L H F Stop T
Stop H Q T F Stop Y R D T Met A E G R C F R L G P Stop V Q R R Q R Stop A V G A G
Stop N V Q R P D S S V C Stop S S D S W Q C L F A A E K R A T H R C W H L Q Met L Y
H H F Stop R Q G E C Stop P Stop V Stop N W S L Q H A G S E C G L Stop C Q L R D L A
V Stop G S P Met V P P A H S G L G I P S Stop P G S Q L L G S L Q Y Q L Stop A E L Stop
E C D H E G C V C A L Q C Y D Q Q H I L L Y D Stop K Stop H C Q S N R G Y Q S D R I
G D Q K A E S P T A A K L K G F S V C L F F L C H Q L G T S A S Q P L P D A K I Met C
L G H K K A C K V I V T T G I Y R T I S P P D Met T Stop F Y I S G R K Stop I H I Stop K
S G V S K Q E Q E T K R S Q K Q K A P I Stop T R Stop I Y L Q R H I R S W E N N S C E L
D K C V K S D K Stop N A R G D K C I P R S Q G P P P A C H L G S E R T G Stop C Met F
F V S E F L V I C A V Met L L Stop G S P W K V Y P N I S T S Y I P Q I K L Stop Y V P
Stop D A A N Stop L P L R N S G A A A F Stop Stop W V K Stop F T F Y D A S K G A L A
S L P N Stop Q Met P K L R K Met I I I L A Stop T E Q S A T P I L Stop I N Stop A P S F
Stop T N K C G F I S Q Met Met F I R E W S R E G P F T L T I W H Y V I T S S E A S P F H
P A W T A K T S V F N S I Stop S S G T Q L G Stop F R P P S P G E C L K T I L V T S Met
R E W R R I Q C Y Y Q L V D K G Q G C C S T S Y H V Q D V S P L Q L P N P K C Q L C
Q D Stop E T L V L S R K G P G K R G A N K S V C F S H Stop S L A N K H S V S L A A A
S A Q R A R T L S G T R I T S L S E Q S Stop Q G L W E Met P D G I I F S L L S F Stop V
S F P S F Y P A S Q V L Stop E K C L S S S S G F L T L N L D L Q T L P G H N S N Stop G
N K H I P S Met K H T Q T F E S K D N D C L N Stop G L E E Stop S F E G K E Y F V S S P
L P T L F Met C Stop P L P S W T L E P R Stop L Y Y Met L L Stop K T D F R V L I V Q E
N D Stop I Y I S

5'3' Frame 2

X E G S G Q L H S A S T Q I R W E P S P A Met A S L G Q I L F W S I I S I I I I L A G A I A
L I I G F G I S G R H S I T V T T V A S A G N I G E D G I L S C T F E P D I K L S D I V I Q
W L K E G V L G L V H E F K E G K D E L S E Q D E Met F R G R T A V F A D Q V I V G
N A S L R L K N V Q L T D A G T Y K C Y I I T S K G K G N A N L E Y K T G A F S Met P
E V N V D Y N A S S E T L R C E A P R W F P Q P T V V W A S Q V D Q G A N F S E V S
N T S F E L N S E N V T Met K V V S V L Y N V T I N N T Y S C Met I E N D I A K A T G
D I K V T E S E I K R R S H L Q L L N S K A S L C V S S F F A I S W A L L P L S P Y L
Met L K Stop C A L A T K K H A K S L L Q Q G S T E L F H H Q I Stop P S F I F L G G N E
F I S R S L E Stop A N K S K K Q K E A K S R R L Q Y E Q D K S I F K D I L E V G K I I H
V N Stop T S V L R V I S K Met H V E T S A S P D L R D L P L P V T W G V R G Q D S A
C S L S L N F Stop L Y V L Stop C C S E E A P G K S I P T Y P H L I F H K L S C S Met Y
P K T L L I D C H F A T Q G R L H F S N G S N D S L F Met Met L P K V P W L L F P T D

FIGURE 2

K C Q S Stop E K Stop S Stop F Stop H K Q S S R R H R F Y K Stop T E H L L F K Q T N A
G L F L R Stop C S S V N G P G K D L S P Stop L Y G I Met S S Q A L R L L L S I L R G Q
L R P Q F S I A S R A V G L S W G D F A P H L R G N V Stop R Q F W L P Q Stop G S G G
G Y S A T T N Stop W I K A R D A A Q P P T Met Y R T S P H Y N Y P I R S V N C V R T
K K P W F Stop V E K G L E R G E P T N L S A S H I S H W Q I S I L S L W L L P Q H R E
P E L Y R A P G Stop H L S V N R V D K A Y G K C L Met G L S S A C Stop A S K F L S L
H S T L Q A K F C K R N A Stop V L A Q V F L L Stop I Stop I S R P F L A T I Q I K A T N
I Y L P Stop S T H R L L K A R T Met T A Stop I E A L R N E A L K E K N T L F P A P F P
H S S C V N H C L P G P W S H G D C I T C C Y R K L I L E F Stop S F K R Met I K Y T F
P

5'3' Frame 3

X K A A G S S T Q P V P R Y A G N L P Q P W L P W G R S S S G A Stop L A S S L F W L
E Q L H S S L A L V F Q G D T P S Q S L L S P Q L G T L G R Met E S Stop A A L L N L T
S N F L I S Stop Y N G Stop R K V F Stop A W S Met S S K K A K Met S C R S R Met K C
S E A G Q Q C L L I K Stop Stop L A Met P L C G Stop K T C N S Q Met L A P T N V I S S
L L K A R G Met L T L S I K L E P S A C R K Stop Met W T I Met P A Q R P C G V R L P D
G S P S P Q W S G H P K L T R E P T S R K S P I P A L S Stop T L R Met Stop P Stop R L C
L C S T Met L R S T T H T P V Stop L K Met T L P K Q Q G I S K Stop Q N R R S K G G V
T Y S C Stop T Q R L L C V S L L S L P S A G H F C L S A L T Stop C Stop N N V P W P Q
K S Met Q S H C Y N R D L Q N Y F T T R Y D L V L Y F W E E Met N S Y L E V W S E Q
T R A R N K K K P K A E G S N Met N K I N L S S K T Y Stop K L G K Stop F Met Stop T
R Q V C Stop E Stop Stop V K C T W R Q V H P Q I S G T S P C L S P G E Stop E D R I V
H V L C L Stop I F S Y Met C C N V A L R K P L E S L S Q H I H I L Y S T N Stop A V V C
T L R R C Stop L T A T S Q L R G G C I L V Met G Q Met I H F L Stop C F Q R C L G F S S
Q L T N A K V E K N D H N F S I N R A V G D T D F I N K L S T F F L N K Q Met R V Y F
S D D V H P Stop Met V Q G R T F H L D Y Met A L C H H K L Stop G F S F P S C V D S
Stop D L S F Q Stop H L E Q W D S A G V I S P P I S G G Met S E D N F G Y L N E G V E
E D T V L L P T S G Stop R P G Met L L N L L P C T G R L P I T T T Q S E V S T V S G L R
N P G F E Stop K R A W K E G S Q Q I C L L L T L V I G K Stop A F C L F G C C L S T E S
Q N S I G H Q D N I S Q Stop T E L T R P Met G N A Stop W D Y L Q L V E L L S F F P F I
L P C K P S S V R E Met P E F Stop L R F S Y S E F R S P D P S W P Q F K L R Q Q T Y T
F H E A H T D F Stop K Q G Q Stop L L E L R P Stop G Met K L Stop R K R I L C F Q P P S
H T L H V L T T A F L D L G A T V T V L H V V I E N Stop F Stop S S D R S R E Stop L N
I H F

3'5' Frame 1

R K C I F N H S L E R S E L Stop N Q F S I T T C N T V T V A P R S R K A V V N T Stop R
V W E G G W K Q S I L F L Q S F I P Q G L N S S S H C P C F Q K S V C A S W K V Y V C
C L N L N C G Q E G S G D L N S E Stop E N L S Stop N S G I S L T E L G L Q G R Met K
G K K L R S S T S Stop R Stop S H Q A F P I G L V N S V H Stop E Met L S W C P I E F W
L S V L R Q Q P K R Q N A Y L P Met T N V R S R Q I C W L P S F Q A L F Y S K P G F L
S P D T V D T S D W V V V Met G R R P V H G R R L S S I P G L Y P L V G S S T V S S S
T P S L R Stop P K L S S D I P P E Met G G E I T P A E S H C S R C Y Stop K L R S Stop L
S T Q D G K E K P Q S L Stop Stop H N A I Stop S R Stop K V L P W T I H G Stop T S S E
K Stop T R I C L F K K K V L S L F I K S V S P T A L F Met L K L Stop S F F S T L A F V S

FIGURE 2

W E E K P R H L W K H H K K Stop I I Stop P I T K Met Q P P L S C E V A V N Stop Q R L
R V H T T A Stop F V E Y K Met W I C W D R L S R G F L R A T L Q H I Stop L K I Q R Q
R T C T I L S S H S P G D R Q G E V P E I W G C T C L H V H F T Y H S Stop H T C L V H
Met N Y F P N F Stop Y V F E D R F I L F I L E P S A F G F F L F L A L V C S L Q T S R Y
E F I S S Q K Y K T R S Y L V V K Stop F C R S L L Stop Q Stop L C Met L F C G Q G T L
F Stop H Q V R A E R Q K C P A D G K E R R D T Q R S L Stop V Stop Q L Stop V T P P F
D L R F C H F D I P C C F G N V I F N H T G V C V V D R N I V E H R H N L H G H I L R
V Q L K A G I G D F R E V G S L V N L G C P D H C G L G E P S G S L T P Q G L Stop A
G I I V H I H F R H A E G S S F I L K V S I P L A F R S D D I T F V G A S I C E L H V F Q P
Q R G I A N Y H L I S K H C C P A S E H F I L L R Q L I F A F F E L Met D Q A Stop N T F
L Q P L Y H D I R K F D V R F K S A A Q D S I L P N V P S Stop G D S S D C D G V S P
Stop N T K A N D E C N C S S Q N N D D A N Y A P E E D L P Q G S H G W G R F P A Y
L G T G Stop V E L P A A F X X

3'5' Frame 2
G N V Y L I I L L N D Q N S K I S F L Stop Q H V I Q S P W L Q G P G R Q W L T H E E C
G K G A G N K V F F S F K A S F L K A S I Q A V I V L A F K S L C V L H G R Y Met F V
A L I Stop I V A R K G L E I Stop I Q S K K T Stop A R T Q A F L L Q N L A C R V E Stop
R E R N L E A Q Q A E D N P I R H F P Stop A L S T L F T E R C Y P G A R Stop S S G S L
C Stop G S S Q R D R Met L I C Q Stop L Met Stop E A D R F V G S P L S R P F S T Q N Q
G F L V L T Q L T L R I G Stop L Stop W G D V L Y Met V G G Stop A A S L A F I H Stop
L V V A L Y P P P L P H Stop G N Q N C L Q T F P R R W G A K S P Q L S P T A L D A I E
N Stop G L S C P R R Met E R R S L R A C D D I Met P Y S Q G E R S F P G P F T D E H H
L R N K P A F V C L K R R C S V Y L Stop N R C R R L L C L C Stop N Y D H F S Q L W
H L S V G K R S Q G T F G S I I K S E S F D P L L K C S R P Stop V A K W Q S I S S V L G
Y I L Q L N L W N I R C G Y V G I D F P G A S S E Q H Y S T Y N Stop K F R D K E H A L
S C P L T P Q V T G R G R S L R S G D A L V S T C I L L I T L N T L V Stop F T Stop I I F
P T S N Met S L K I D L S C S Y W S L L L L A S F C F L L L F A H S R L L D Met N S F P
P R N I K L G H I W W Stop N S S V D P C C N N D F A C F F V A K A H Y F S I R Stop G
L R G R S A Q L Met A K K E E T H R E A F E F S S C R Stop L R L L I S D S V T L I S P V
A L A Met S F S I I Q E Y V L L I V T L Stop S T D T T F Met V T F S E F S S K L V L E T
S E K L A P W S T W D A Q T T V G W G N H R G A S H R K V S E L A L Stop S T F T S G
Met L K A P V L Y S R L A F P L P L E V Met I Stop H L Stop V P A S V S C T F F S R K E
A L P T I T Stop S A N T A V R P L N I S S C S D S S S L P S L N S W T K P K T P S F S H
C I T I S E S L Met S G S K V Q L R I P S S P Met F P A E A T V V T V Met E C L P E I P K
P Met Met S A I A P A R I Met Met Met L I Met L Q K R I C P R E A Met A G E G S Q R I
W V L A E W S C P L P S X

3'5' Frame 3
E Met Y I Stop S F S Stop T I R T L K S V F Y N N Met Stop Y S H R G S K V Q E G S G
Stop H Met K S V G R G L E T K Y S F P S K L H S S R P Q F K Q S L S L L S K V C V C F
Met E G I C L L P Stop F E L W P G R V W R S K F R V R K P E L E L R H F S Y R T W L
A G Stop N E G K E T Stop K L N K L K I I P S G I S H R P C Q L C S L R D V I L V P D R
V L A L C A E A A A K E T E C L F A N D Stop C E K Q T D L L A P L F P G P F L L K T R
V S Stop S Stop H S Stop H F G L G S C N G E T S C T W Stop E V E Q H P W P L S T S W
Stop Stop H C I L L H S L I E V T K I V F R H S P G D G G R N H P S Stop V P L L Stop

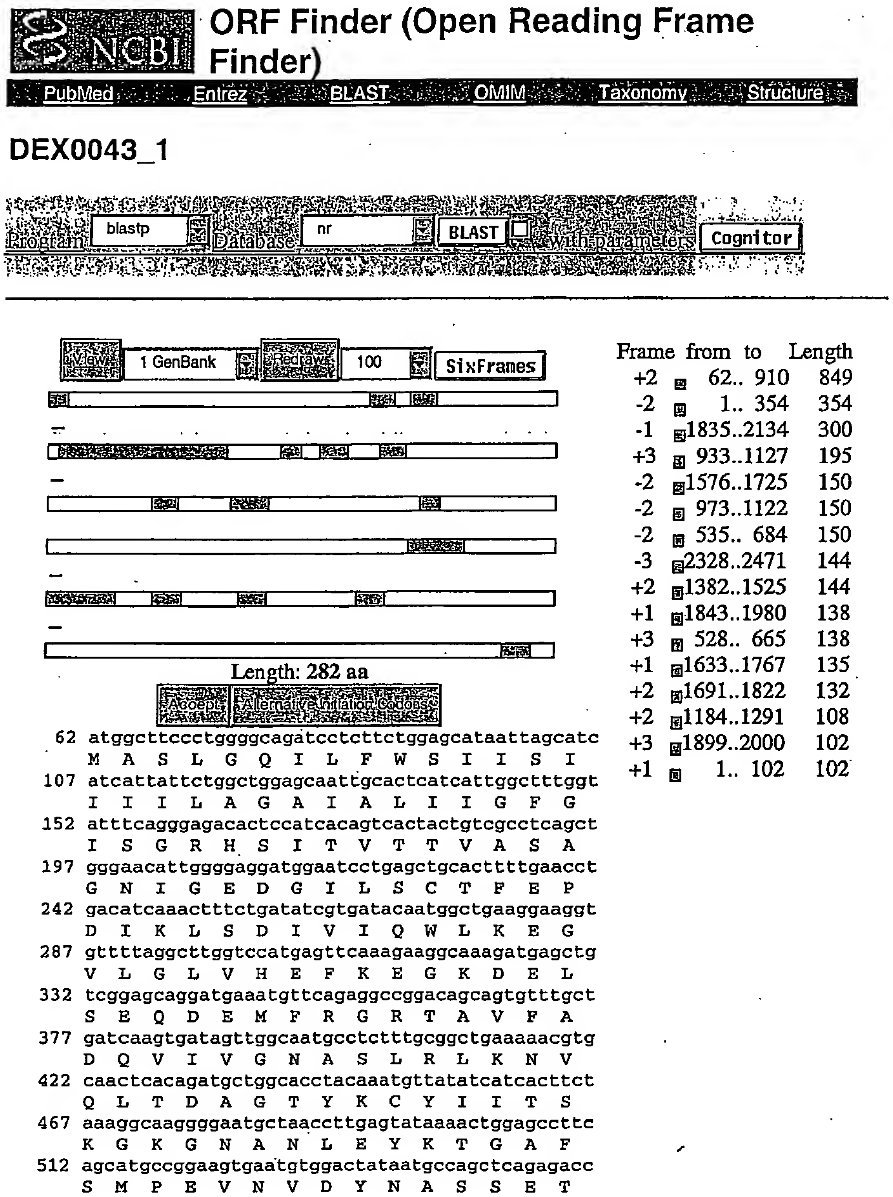
FIGURE 2

4 / 4

Met L L K T E V L A V H A G W K G E A S E L V Met T Stop C H I V K V K G P S L D H S
R Met N I I Stop E I N P H L F V Stop K E G A Q F I Y K I G V A D C S V Y A K I Met I I F
L N F G I C Q L G R E A K A P L E A S Stop K V N H L T H Y Stop N A A A P E L R S G S
Q L A A S Stop G T Y Y S L I C G I Stop D V D Met L G Stop T F Q G L P Q S N I T A H I T
K N S E T K N Met H Y P V L S L P R Stop Q A G G G P Stop D L G Met H L S P R A F Y L
S L L T H L S S S H E L F S Q L L I C L Stop R Stop I Y L V H I G A F C F W L L F V S C S
C L L T P D F Stop I Stop I H F L P E I Stop N Stop V I S G G E I V L Stop I P V V T Met T
L H A F L W P R H I I L A S G K G Stop E A E V P S Stop W Q R K K R H T E K P L S L A
A V G D S A F Stop S P I L S L Stop Y P L L L W Q C H F Q S Y R S Met C C Stop S Stop
H C R A Q T Q P S W S H S Q S S A Q S W Y W R L P R S W L P G Q L G Met P R P L W A
G G T I G E P H T A R S L S W H Y S P H S L P A C Stop R L Q F Y T Q G Stop H S P C L
Stop K Stop Stop Y N I C R C Q H L Stop V A R F S A A K R H C Q L S L D Q Q T L L S G
L Stop T F H P A P T A H L C L L Stop T H G P S L K H L P S A I V S R Y Q K V Stop C Q
V Q K C S S G F H P P Q C S Q L R R Q Stop Stop L Stop W S V S L K Y Q S Q Stop Stop
V Q L L Q P E Stop Stop Stop C Stop L C S R R G S A P G K P W L G K V P S V S G Y W L
S G A A R C L X X

FIGURE 3

1 / 2



ORF Finder (Open Reading Frame Finder)

DEX0043_1

Frame	from	to	Length
+2	62..	910	849
-2	1..	354	354
-1	1835..	2134	300
+3	933..	1127	195
-2	1576..	1725	150
-2	973..	1122	150
-2	535..	684	150
-3	2328..	2471	144
+2	1382..	1525	144
+1	1843..	1980	138
+3	528..	665	138
+1	1633..	1767	135
+2	1691..	1822	132
+2	1184..	1291	108
+3	1899..	2000	102
+1	1..	102	102

Length: 282 aa

```
 62 atggcttccctggggcagatcctcttctggagcataattagcatc
    M  A  S  L  G  Q  I  L  F  W  S  I  I  S  I
107 atcattattctggctggagcaattgcactcatcattggctttggt
    I  I  I  L  A  G  A  I  A  L  I  I  G  F  G
152 atttcagggagacactccatcacagtcactactgtcgcctcagct
    I  S  G  R  H  S  I  T  V  T  T  V  A  S  A
197 gggaacattggggaggatggaatcctgagctgcacttttgaacct
    G  N  I  G  E  D  G  I  L  S  C  T  F  E  P
242 gacatcaaactttctgatatcgtgatacaatggctgaaggaaggt
    D  I  K  L  S  D  I  V  I  Q  W  L  K  E  G
287 gttttaggcttggtccatgagttcaaagaaggcaaagatgagctg
    V  L  G  L  V  H  E  F  K  E  G  K  D  E  L
332 tcggagcaggatgaaatgttcagaggccggacagcagtgtttgct
    S  E  Q  D  E  M  F  R  G  R  T  A  V  F  A
377 gatcaagtgatagttggcaatgcctctttgcggctgaaaaacgtg
    D  Q  V  I  V  G  N  A  S  L  R  L  K  N  V
422 caactcacagatgctggcacctacaaatgttatatcatcacttct
    Q  L  T  D  A  G  T  Y  K  C  Y  I  I  T  S
467 aaaggcaaggggaatgctaaccttgagtataaaactggagccttc
    K  G  K  G  N  A  N  L  E  Y  K  T  G  A  F
512 agcatgccggaagtgaatgtggactataatgccagctcagagacc
    S  M  P  E  V  N  V  D  Y  N  A  S  S  E  T
```

FIGURE 3

2 / 2

```
557 ttgcggtgtgaggctccccgatggttcccccagcccacagtggtc
    L  R  C  E  A  P  R  W  F  P  Q  P  T  V  V
602 tgggcatcccaagttgaccagggagccaacttctcggaagtctcc
    W  A  S  Q  V  D  Q  G  A  N  F  S  E  V  S
647 aataccagctttgagctgaactctgagaatgtgaccatgaaggtt
    N  T  S  F  E  L  N  S  E  N  V  T  M  K  V
692 gtgtctgtgctctacaatgttacgatcaacaacacatactcctgt
    V  S  V  L  Y  N  V  T  I  N  N  T  Y  S  C
737 atgattgaaaatgacattgccaaagcaacaggggatatcaaagtg
    M  I  E  N  D  I  A  K  A  T  G  D  I  K  V
782 acagaatcggagatcaaaaggcggagtcacctacagctgctaaac
    T  E  S  E  I  K  R  R  S  H  L  Q  L  L  N
827 tcaaaggcttctctgtgtgtctcttctttctttgccatcagctgg
    S  K  A  S  L  C  V  S  S  F  F  A  I  S  W
872 gcacttctgcctctcagcccttacctgatgctaaaataa 910
    A  L  L  P  L  S  P  Y  L  M  L  K  *
```

ExPASy Home page Site Map Search ExPASy Contact us

Search ExPASy web site for translate Go Clear

The ExPASy Molecular Biology Server

History of changes, improvements and new features

If you subscribe to our Swiss-Flash service of electronic bulletins, you can receive these and other news by electronic mail.

October 14, 2004

- *Tools*
  Aldente is a tool to identify proteins from peptide mass fingerprinting data. This new, fast and powerful PMF tool uses the Hough transform to determine the mass spectrometer deviation, to realign the experimental masses and to exclude outliers (More information).

- *Mirror site*
  We are happy to announce a new **ExPASy mirror site in Brazil**, http://br.expasy.org/, hosted by the Laboratório Nacional de Computação Científica, Petropolis

June 4, 2004

- The Melanie page has been restyled. It has been redesigned by the occasion of the announcement of SIB, Genebio and Amersham Biosciences joining forces to create one single 2D image analysis. Melanie was chosen to be integrated into ImageMaster™ 2D Platinum software.

April 14, 2004

- *UniProt*
  Since the last Swiss-Flash Bulletin, the universal protein resource, UniProt has been released publically. Many ExPASy pages and services have changed to accommodate different aspects of the UniProt knowledgebase and UniRef, the non-redundant reference databases of UniProt.

  In particular, the ExPASy BLAST interface now allows to launch a sequence similarity search against the UniRef clusters UniRef100, UniRef90 and UniRef50.

  Implicit links to UniRef50 and UniRef90 have been added to the NiceProt view of UniProt knowledgebase entries.

- *FTP server structure*
  As announced in a previous Swiss-Flash bulletin, the structure of the ExPASy ftp server has changed. Please refer to this previous announcement for details.

- *Swiss-Prot/TrEMBL (UniProt knowledgebase)*
  A note to Swiss-Prot and TrEMBL users: Please note that we have a long list of planned format changes to be introduced in the next few months.

  In the NiceProt view for Swiss-Prot/TrEMBL entries we have added implicit links to the Swiss-Model repository of 3D homology models (SMR).

  It is now possible to submit all splice isoforms annotated in one Swiss-Prot entry to a multiple alignment, or to retrieve the . quences of all these isoforms, e.g. from http://www.expasy.org/cgi-bin/niceprot.pl?P29590#comments or from http://www.expasy.org/cgi-bin/get-varsplic.pl?P29590-4

- *PROSITE*
  PSview The view of PROSITE documentation entries contains new functionalities. When a 3D structure is described in the text, a direct link to a 3D image of the domain is provided. The Swiss-Prot match list of each signature can be visualized as a multiple alignment, or as a taxonomic distribution graph. For PROSITE profiles, a domain arrangement view is also provided where active sites and disulfide bridges annotated in Swiss-Prot entries are superimposed on PROSITE domains. see the following links for more details: http://www.expasy.org/cgi-bin/nicedoc.pl?PDOC50119 http://www.expasy.org/cgi-bin/prosite/PSView.cgi?ac=PS50119&onebyarch=1&trembl=1&hscale=0.6

- *ENZYME*
  Access to ENZYME entries by class, subclass etc. has been improved. It is now possible to easily retrieve all ENZYME or Swiss-Prot entries corresponding to a given ENZYME class. This functionality is available from a given ENZYME entry or for a given ENZYME class.

The legends for the Biochemical Pathways have been made available in html and pdf format.

- *Tools*
  Myristoylator is a new tool designed to predict N-terminal myristoylation of proteins by neural networks. N-terminal myristoylation is a post-translational modification that causes the addition of a myristate group to the N-terminal glycine of an amino acid chain.

September 26, 2003

- *Swiss-Prot variant web pages*
  Missense mutation leading to single amino acid polymorphism (SAP) is the type of mutation most frequently related to human diseases. We have created Swiss-Prot Variant web pages to provide a summary of available sequence information as well as additional structural information on each variant. In particular, wherever possible, SAPs are modeled onto 3D protein structures and the users can visualize the models. The 3D models are updated with each weekly Swiss-Prot release. The Swiss-Prot variant pages are accessible from the NiceProt view of a Swiss-Prot entry (e.g. P06737) on the ExPASy server, via a hyperlink created for the stable and unique identifier FTId of each human SAP (e.g. VAR_007908).

- *Recent and forthcoming changes in Swiss-Prot*
  With Swiss-Prot release 41, we have introduced two documents to announce recent and forthcoming format changes in Swiss-Prot and TrEMBL.These documents replace the corresponding sections of the release notes, and contain detailed information about new keywords, new feature keys and comment topics, new cross-references and other format changes. Explicit links to new databases will no longer be announced here (i.e. under "What's new on ExPASy"), but in the document "Recent changes".

- *Implicit links*
  Implicit links (i.e. added on the fly to Swiss-Prot/TrEMBL entries when viewed with NiceProt) to the following databases have been added recently:
  - GenAtlas - A human gene database, e.g. P04406
  - HOBACGEN - Homologous bacterial genes database, e.g. P02937
  - HOVERGEN - Homologous vertebrate genes database, e.g. P02304
  - TAIR - The Arabidopsis Information Resource, e.g. Q38828
  - WorfDB - The C.elegans ORFeome cloning project, e.g. Q17330

- WormBase - ..tabase on genetics, genomics and biology ..f C. elegans, e.g. Q17330

Information about the criteria for the creation of links to each of these databases can be found in the Swiss-Prot document List of databases cross-referenced in Swiss-Prot.

Whenever reference is made to the Transport Commission (TC) System in Swiss-Prot comments lines (CC), a link is added from the NiceProt view to the Transport Protein Database (example: P37905).

- *Visualization tool for peptide mass fingerprinting identification results*
  We have installed Biograph Applet v2.0, intended for the visualization of results of the PeptIdent, FindPept and FindMod tools. Links to Biograph are available as part of PeptIdent, FindPept and FindMod result pages.

March 21, 2003

New cross-references have been introduced in Swiss-Prot:

- Explicit links to GeneDb SPombe, the Schizosaccharomyces pombe GeneDB, example: O94534.
- Implicit links to CleanEx, a gateway to public gene expression data via officially approved gene names, example: P02751.

There is a new Swiss-Prot document, arath.txt - Index of Arabidopsis thaliana entries and their corresponding gene designations.

An interface to PRATT has been implemented on ExPASy. PRATT is a tool to discover patterns that are conserved in a set of protein sequences. The patterns are reported using the PROSITE format. The ExPASy BLAST result representation has been modified to allow direct submission of a number of sequences to PRATT.

The ExPASy BLAST interface now allows to perform tblastn searches against individual microbial genomes (EMBL genome records, including plasmids).

Throughout the ExPASy server, the navigation bar at the top of every page now includes a search bar, for quick access to Swiss-Prot, TrEMBL, PROSITE, SWISS-2DPAGE, ENZYME, Taxonomy, HAMAP and ExPASy site search.

November 19, 2002

We are happy to announce a new **ExPASy mirror site in Bolivia**, http://bo.expasy.org, hosted by the Universidad Católica Boliviana in Cochabamba.

October 25, 2002

ExPASyBar, a very useful navigation bar to the most important databases and tools on the ExPASy server, has been developed by Martin Hassman, with input from the ExPASy team. ExPASyBar is an add-on to the Mozilla web browser (i.e. it does not work with Netscape, MS Internet Explorer and other browsers). Installation is very simple. ExPASyBar can be configured to use the ExPASy mirror of the user's choice (in the Edit/Preferences/Advanced/ExPASyBar menu of Mozilla).

August 27, 2002

The last "What's new on ExPASy" is more than a year old, which means that some of the changes and "new" features and services are not all that new anymore.... We are trying to list here the most important changes, and we will try to report new tools and documents again more frequently in the future!

== ExPASy ==

- We are happy to announce that since the beginning of the year 2002, there is an **ExPASy mirror site in the USA**, http://us.expasy.org, hosted by the North Carolina Supercomputing Center (NCSC). Some users may have noticed upon their connection to www.expasy.org, that they are redirected to a mirror site that is closer to their geographic location, or that is less heavily loaded. If you feel that you are redirected to a mirror site for which the network connection is slow, please let us know .
- News on the FTP server:
  1. **PROSITE updated data** and documentation files are now made available via FTP even between releases, in the directory /databases/prosite/release_with_updates/. This data always corresponds to the version of the database available for web access via the PROSITE page.
  2. Up-to-date plain-text versions of all **SWISS-PROT documents** can be downloaded by ftp, in the directory /databases/swiss-prot/updated_doc/.
  3. Three "special selections" have been added:
     - merops.seq - all SWISS-PROT entries cross-referenced to the MEROPS database
     - mitoch.seq - Mitochondrion encoded proteins (entries with "Mitochondrion" on OG lines)
     - plastid.seq - Chloroplast and cyanelle encoded proteins (entries with "Chloroplast" or "Cyanelle" on OG lines)

== TOOLS ==

Two tools have been added to our collection of sequence analysis and proteomics tools:

- The Sulfinator predicts **tyrosine sulfation sites** in protein sequences, based on Hidden Markov Models.
- PeptideCutter predicts potential **cleavage sites** cleaved by proteases or chemicals in a given protein sequence. It displays the query sequence with the possible cleavage sites mapped on it, as well as a table of cleavage site positions.

Major updates have been performed on two tools:

- The **ScanProsite** interface has been remodeled, with more options and databases, and a graphical view of the results. A standalone program, ps_scan is now available.
- The **BLAST** interface now allows searches in PDB. The output page displays hits found with Pfam HMMs and PROSITE profiles on the query sequence.

== SWISS-PROT ==

- New **cross-references** have been *introduced* to various databases: AraC-XylS, Genew, Gramene, several 2D-PAGE databases, Ensembl, GeneLynx, ListiList, ModBase, PhosSite, ProtoNet, Source and TIGRFAMs.
  The List of databases cross-referenced in SWISS-PROT contains, for each database, a short description, the link type (explicit or implicit), and the server URL. In the case of explicit links, you can click on the word "Explicit" (example: Genew) to retrieve a list of all SWISS-PROT entries linked to the corresponding database.
  The cross-references to the following databases have been *deleted*, because the databases are either no longer available on the WWW, or because they have become commercial even for academic users: CarbBank, DOMO, GCRDb, Mendel, YEPD and YPD.
- The NiceProt view of SWISS-PROT has been further improved: access to documentation has been facilitated by adding "mouse-over" hypertext links from various sections in NiceProt to

the corresponding information in the user manual. Those hypertext links, which give access to documentation rather than the data related to the protein entry, are visually different from the ordinary hyperlinks. While they are not immediately recognizable as such, the user can see that they are clickable by moving the mouse pointer over the section headings such as "References" or "Keywords". A short description of the linked information appears at the bottom of the web browser, and when clicked, a small additional window is opened with related information extracted from the user manual.

Similarly, in the "Cross-references" section, the names of the databases to which an entry is cross-referenced are linked to the corresponding sections in the document dbxref.txt (List of databases cross-referenced in SWISS-PROT).

- Three SWISS-PROT documents have been released since the last announcement:
  - bucai.txt - Index of Buchnera aphidicola (subsp. Acyrthosiphon pisum) entries
  - mycpn.txt - Index of Mycoplasma pneumoniae strain M129 entries
  - plasmid.txt - List of plasmids
- The Human proteomics initiative (HPI) status report page has been remodeled and now contains more detailed information about the status of annotation of human SWISS-PROT entries.
- The HAMAP project aims to annotate semi-automatically **complete bacterial and archaeal proteomes** up to the quality standards of SWISS-PROT. Several proteomes have already been completed and are continuously updated. Up-to-date statistics are available on the HAMAP status page
- Note that upcoming format changes in the next SWISS-PROT release are always described in the release notes for the current release.
- Although not hosted physically on the ExPASy server, the NEWT Taxonomy browser is provided and maintained by members of the SWISS-PROT group, and serves as an entry point into SWISS-PROT and TrEMBL using taxonomic search criteria.

== SWISS-2DPAGE ==

New **cross-references, reference maps,** and a **document** have been added:

- Cross-references to recent fully federated 2-DE databases, built with the Make2ddb package, are provided. These are now COMPLUYEAST-2DPAGE, PHCI-2DPAGE, PMMA-2DPAGE, and Siena-2DPAGE. The list of links is updated each time a SWISS-2DPAGE release is completed.
- SDS-PAGE and 2-D PAGE of nucleolar proteins from Human HeLa cells have been added to the list of reference maps . These masters are named respectively NUCLEOLI_HELA_1D_HUMAN and NUCLEOLI_HELA_2D_HUMAN.
- A FAQ (Frequently Asked Questions) has been provided. We hope you will find answers to most of your questions in this new document.

June 30, 2001

New cross-references have been added from relevant SWISS-PROT entries to three databases:

- SMART - Protein domain database (example: O43707).
- Leproma - Database dedicated to the analysis of the genome of the leprosy bacillus Mycobacterium leprae (example: Q9CBW4).
- MypuList - Mycoplasma pulmonis genome database (example: P58174).

The keyword "Complete proteome" has been introduced to all SWISS-PROT/TrEMBL entries describing a protein which is thought to be expressed by an organism whose genome has been completely sequenced. This keyword is so far only used for microbial (bacterial and archaeal) proteins. A complete set of proteins from a microbial genome can therefore be obtained using this keyword across SWISS-PROT and TrEMBL.

A new and improved version of the NiceProt view of SWISS-PROT is available (example). Some of its new features are:

- It provides a link to a printer-friendly view of a SWISS-PROT entry.
- It displays the length of certain features in the FT lines.
- It provides access to a new tool, the 'Feature aligner' which allows to select features for submission to the ClustalW multiple alignment program.

SWISS-PROT release statistics are now available for every update of the database. Among other parameters, statistics about database growth, average sequence lengths and amino acid composition, taxonomic origin, journal citations and database cross-references are presented, including some graphics.

A new view is available within the SRS Sequence Retrieval System. It displays, for each protein corresponding to a user query, gene name(s) and organism (in addition to the parameters ID, AC, description and sequence length which are displayed by the default view "Short description"). This new view is entitled "**Long description**" and is available from the menu "Use view ..." in the SRS query form.

The SIB Blast interface (accessible also via "Quick BLAST" or from the bottom of every SWISS-PROT/TrEMBL entry) now offers the possibility to restrict the similarity search by using taxonomic criteria. A "Taxonomic View" of the results can also be obtained via the BLAST result page.

L'équipe Swiss-Prot a le plaisir de vous présenter le premier article de "Protéines à la Une", sa nouvelle rubrique de vulgarisation scientifique dédiée aux protéines qui font parler d'elles dans l'actualité.

January 18, 2001

- **SWISS-PROT**
  New cross-references have been added to three additional databases:
  - GlycoSuiteDB - a database of glycan structures; explicit links
    example: P00750
  - GeneCensus - a compilation of ORF data for the Saccharomyces genome; implicit links
    example: Q01802
  - HUGE - a database of human unidentified gene-encoded large proteins; implicit links
    example: P42330

- **NiceProt & SIB BLAST** The NiceProt view of SWISS-PROT/TrEMBL entries now contains a direct submission button requesting a blastp homology search of the protein against SWISS-PROT/TrEMBL/TrEMBLnew, on the SIB BLAST server (*"Quick BlastP search"*). In the results of SIB BLAST searches on ExPASy (normal or "NiceBlast" output formats), the user can select a number of matching sequences and directly submit them to a ClustalW search, or retrieve and download the corresponding SWISS-PROT/TrEMBL entries.

- **Proteomics tools**
  - FindPept: This new tool can identify peptides that result from unspecific cleavage of proteins from their experimental masses, taking into account artefactual chemical modifications,

post-translational modifications (PTM) and protease autolytic cleavage.

- PeptIdent: Several new features have been added.
    - When searching SWISS-PROT, all alternative splice isoforms described in SWISS-PROT feature tables are included in the search (e.g. Isoform 12S of O43184).
    - New organism classes can be searched. For each of the available taxonomic available (e.g. Mammalia), a new section (e.g. other Mammalia) has been added, which comprises all entries not corresponding to any of the searchable subclasses (e.g. all Mammalia except human, bovine, rabbit, and other rodents).
    - For each matching protein in a PeptIdent result, buttons are available which allow further analysis of the protein by direct submission of the data to FindMod, FindPept, GlycoMod, PeptideMass and BioGraph.
- GlycoMod: Possible oligosaccharide structures suggested by GlycoMod are linked to the GlycoSuiteDB database of glycan structures, if they are reported in this database. The user can also select to display compositions reported in GlycoSuiteDB separately from the compositions not known in the database.

October 28, 2000

Several new features have been implemented on ExPASy during the last few months:

- The Swiss Center for Scientific Computing (CSCS) and the Swiss Institute of Bioinformatics provide a powerful and rapid new BLAST server. A submission form to this server is available from the bottom of each SWISS- PROT/TrEMBL entry on ExPASy. Results of blastp similarity searches submitted from this form are now parsed and displayed in a more user-friendly way, including a graphical representation and a link to NiceBlast. NiceBlast is a html table detailing complete descriptions of all matching proteins, including the full protein name, gene name, sequence length and organism.

- Sequences of alternatively spliced isoforms of the same protein are documented in the feature table of that protein sequence record. In collaboration with the SWISS-PROT group at EBI, a program varsplic.pl has been written to generate additional records from SWISS-PROT and TrEMBL, one for each splice isoform of each protein. The resulting data sets for SWISS-PROT and TrEMBL are available on our ftp server, along with a more detailed description of the project and information on how to obtain a local copy of the varsplic.pl program.

  The additional isoform entries have been added to the SWISS- PROT/TrEMBL databases underlying the BLAST server at SIB/CSCS Switzerland, and ScanProsite. Gradually, all other tools on ExPASy will be modified to handle splice isoforms. The NiceProt view of SWISS-PROT/TrEMBL provides links from the isoform name in the feature table (example: Q01432) to a page displaying the sequence of the corresponding isoform.

- In the framework of the HAMAP project, we provide clean non-redundant SWISS-PROT/TrEMBL data sets for all completely sequenced microbial genomes. These files are available on the ExPASy ftp server in SWISS-PROT and Fasta format, and can also be used for similarity searches on the SIB Blast server ("microbial proteomes").

  A **Genomic Proximity Viewer** is available for those microbial genomes where an ORF numbering system exists. For those organisms, it is possible to click on the ORF name in the SWISS-PROT/TrEMBL GN (gene) lines to obtain a list of proteins encoded by genes in proximity (example: P46448). The tool is also accessible from the HAMAP complete proteome pages of those organisms. Example: Borrelia burgdorferi.

- The following cross-references have been added to relevant SWISS-PROT/TrEMBL entries:
  - InterPro - the Integrated Resource of Protein Families, Domains and Sites, integrating PROSITE, Pfam, PRINTS and ProDom. A link to a graphical view of the domain structure is also available; example: O15197.
  - MEROPS - a peptidase database; example: O96009.
  - NucleaRDB - a database of nuclear receptors (implicit links); example: O09018.
  - DIP - Database of Interacting Proteins (implicit links); example: P10275
- The Compute pI/Mw tool, if called for a list of proteins, can now produce, in addition to the usual verbose format, a table in text format that can be exported to an external application.
- Protein Spotlight is a periodical electronic review from the SWISS-PROT group. It is published on a monthly basis and consists of articles focused on particular proteins of interest. You can subscribe to receive each issue, free of charge, in HTML or PDF format.

April 26, 2000

Proteomics Tools:

- We are happy to announce a new tool in our suite of ExPASy protein identification and characterization tools:
  GlycoMod is a tool that can predict the possible oligosaccharide structures occurring on proteins from their experimentally determined masses. The program can be used for free or derivatized oligosaccharides and for glycopeptides. GlycoMod has been developed in collaboration with Nicolle Packer, initially at Macquarie University, Sydney, and later at Proteome Systems Ltd. GlycanMass is an associated tool which allows to calculate the mass of an oligosaccharide structure from its oligosaccharide composition.
- Detailed documentation is now available for the PeptIdent peptide mass fingerprinting identification tool.
- A number of new functionalities have been added to FindMod:
  - Results can now be obtained by email (as an alternative to receiving them on-line in the browser window), in form of an html file, with exactly the same functionality as for on-line display.
  - Several new enzymes have been added, mainly different versions of Chymotrypsin.
  - Results given in the "potential amino acid substitutions" table have been refined:
    - We no longer suggest amino acid (aa) substitutions occurring on the enzyme cleavage site and substituting the aa for an aa at which the enzyme does not cleave.
    - If the suggested aa substitution corresponds to a sequence variant or conflict as annotated in the SWISS-PROT feature table, this substitution is highlighted in color (green background for that table line), and a hypertext link is provided to the corresponding annotated variant or conflict.
- Compute pI/Mw can now be used with a file uploaded from the user's computer, if this file contains a list of SWISS-PROT/TrEMBL IDs/ACs.

SWISS-PROT:

- Dotlet, a diagonal dot-matrix program drawing a dotplot of two sequences, has been included in the set of tools that can be called directly from the bottom of each SWISS-PROT/TrEMBL entry on ExPASy. This allows to find repeats within the sequence.

- In the last few months, cross-references to the following database. have been added to relevant SWISS-PROT entries:
  - TubercuList - for entries from Mycobacterium tuberculosis
  - PRINTS - Protein fingerprint database
  - implicit links to BLOCKS - a database of multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins.

  Example entry: Q50705.
- There are 6 new SWISS-PROT documents:
  - humchr08.txt: Index of protein sequence entries encoded on human chromosome 8.
  - humchr09.txt:Index of protein sequence entries encoded on human chromosome 9.
  - humchr10.txt:Index of protein sequence entries encoded on human chromosome 10.
  - humchr11.txt:Index of protein sequence entries encoded on human chromosome 11.
  - dbxref.txt: List of databases cross-referenced in SWISS-PROT.
  - rprowaze.txt: Index of Rickettsia prowazekii strain Madrid E entries.

ExPASy:

- We are happy to announce a **new ExPASy mirror site**, at Peking University, China: http://expasy.pku.edu.cn/.
- We have completely revised the ExPASy server access statistics, which were previously frequently incomplete and erroneous. Every month, a table is updated which lists monthly access statistics for the main Swiss ExPASy server and for all our mirror sites.

October 4, 1999

- The ExPASy server has a **new mirror site** for North America, at the Canadian Bioinformatics Resource in Halifax, Canada. It can be reached at the URL http://expasy.cbr.nrc.ca/.
- The SWISS-PROT search by description tool has been extended to TrEMBL.
- There are five new SWISS-PROT documents:
  - humchr12.txt: an index of protein sequence entries encoded on human chromosome 12.
  - humchr14.txt: an index of protein sequence entries encoded on human chromosome 14.
  - humchr15.txt: an index of protein sequence entries encoded on human chromosome 15.
  - humchr16.txt: an index of protein sequence entries encoded on human chromosome 16.
  - annbioch.txt: SWISS-PROT annotation: how is biochemical information assigned to sequence entries
- When scanning a pattern against the SWISS-PROT/TrEMBL databases using the ScanProsite tool, users can now **restrict their searches** to an organism or a taxonomic range.
- The NiceSite view of PROSITE (example: PS00101) has been modified to include two new statistical values in its section of numerical results, namely
  Precision (true hits / (true hits + false positives)) and
  Recall (true hits / (true hits + false negatives)).
- A new parameter has been added to the list of parameters computed by the ProtParam tool: The program now calculates the **atomic composition** of a protein, in addition to molecular weight, theoretical pI, amino acid composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY).

June 16, 1999

The 'Nice' view tools for the databases provided on ExPASy (SWISS-PROT, SWISS-2DPAGE, PROSITE, ENZYME) have been developed in order to provide users with an easily readable alternative to the original text file representation.

The following tools are available:

Database	Tool	Example
SWISS-PROT	NiceProt	http://www.expasy.ch/cgi-bin/niceprot.pl?P14060
SWISS-2DPAGE	Nice2DPage	http://www.expasy.ch/cgi-bin/nice2dpage.pl?P00938
PROSITE	NiceSite	http://www.expasy.ch/cgi-bin/nicesite.pl?PS00661 http://www.expasy.ch/cgi-bin/nicedoc.pl?PDOC00566
ENZYME	NiceZyme	http://www.expasy.ch/cgi-bin/nicezyme.pl?2.4.1.1

We have now changed all our tools and database search programs on ExPASy to display the 'Nice' version of a database entry by default. The programs displaying database entries in their original text formats continue to be maintained, and links are available from the 'Nice' views to the corresponding get-xxx-entry programs (e.g. get-sprot-entry).

If you maintain pages with links to entries from the above-mentioned databases, you might be interested to update these links to use the 'Nice' View if you prefer this representation to the original format. Otherwise you are, of course, completely free to keep the get-xxx-entry links.

May 24, 1999

- *Linking to ExPASy*
  We have revised the ExPASy file and directory structure, in order to have the vast amount of data that has accumulated on ExPASy since September 1993 available in a more structured manner, and to facilitate replication on our mirror sites. This has caused certain changes in html links, and we would like to ask our users to update their bookmarks and links accordingly. If in doubt, please refer to the document 'How to create html links to ExPASy'. At the same time we wish to reiterate our announcement of the ExPASy mirror sites in Taiwan and Australia. For your own convenience, please use the mirror site closest to you. Regular users might also bookmark the addresses of all ExPASy mirror sites to use as backup for the rare cases that their favourite ExPASy site is down or unreachable due to network problems.

  Please make sure to update all pointers using the old domain expasy.hcuge.ch, which was replaced by
  http://www.expasy.ch/ in March 1997 (!). The 'expasy.hcuge.ch' address might be disabled in the near future.

- *Protein identification tools*
  AACompIdent and MultiIdent have been revised, and the database choice has been extended to include TrEMBL. Results are now sent to the user in html format (rather than text only), and html links allow direct access to the matching SWISS-PROT/ TrEMBL entries.

- *SWISS-PROT cross-references*
  SWISS-PROT entries from Escherichia Coli entries with 'DR ECOGENE' lines are now directly linked to EcoGene at the University of Miami.
  There is a new type of cross-reference lines for sequence entries from Brachydanio rerio (Zebrafish): these entries are now linked to the Zebrafish Information Network (ZFIN) at the University of Oregon.

- *New features have been added to improve interactivity in accessing SWISS-2DPAGE:*

- All searching functions in the database can be accessed from the top page and results page of each keyword search function (example: search by description). This feature has been designed to facilitate the navigation between the different ways to query the database (by description, by access number, by authors, by full text search).
- A new tool is provided to retrieve in a table all the protein entries identified on a given reference map, with all 2-DE information: spot serial number, pI, Mw, mapping procedure, references (example).
- A new way to query the database is provided. From a user-entered amino acids sequence, one can display the estimated location on a choosen reference map (example).

February 26, 1999

- *Several new features have been added to the PeptIdent peptide mass fingerprinting identification tool:*
  - It is now possible to search SWISS-PROT and/or **TrEMBL**.
  - In the page displaying the PeptIdent results, a button allows to perform a **new search** with slightly modified parameters by giving access to the PeptIdent form filled in with all previously used parameters.
  - For each matching protein, a direct link to BioGraph gives access to a **graphical representation** of the results of the PeptIdent query. BioGraph was developed by Daniel Doubrovkine and Anton Soudovtsev as a student project in the scope of the Bioinformatics course given at Geneva University.
  - The **sequence portion covered by the matching peptides** can optionally be displayed and highlighted in colour, as well as the difference between pI and Mw values of the matching proteins and the user-specified values.

- In the results of the SIM binary sequence alignment tool, a direct link has been addded to the PRSS program from EMBnet-CH which **evaluates the significance of a protein sequence similarity score**.

- Direct links have been added from the comments (CC) lines of relevant SWISS-PROT entries to the SWISS-PROT documents listing ribosomal protein families (e.g. RL2_ECOLI), aminoacyl-tRNA synthetases (e.g. SYC_HUMAN) and 7-transmembrane G-linked receptors (e.g. AA3R_MOUSE).

- Since the introduction of organism classification (OC) terms of the NCBI taxonomy with SWISS-PROT release 37, OS (organism species) lines have been linked to the corresponding pages of the NCBI taxonomy browser.

- The PROSITE full text search tool has been improved. Like in the SWISS-PROT/TrEMBL full text search program, wildcards can be used in query strings and search keywords can be combined with boolean operators.

- We have developed **Nice2DPage**, a tool that provides a user-friendly tabular view of SWISS-2DPAGE entries (example). The 'Nice2DPage View of SWISS-2DPAGE' is accessible from the top of each SWISS-2DPAGE entry on ExPASy.
- New hypertext cross-references have been added to SWISS-2DPAGE entries (e.g. P02997):
  - from the 2D comments lines (MAPPING, EXPRESSION LEVEL...), direct links have been added to the concerned citation in the SWISS-2DPAGE entry
  - from the 2D lines concerning AMINO ACID COMPOSITION and PEPTIDE MASSES data, direct links have been added to the concerned section in the user manual describing data format and protocols.

- Links to the Brenda enzyme database have been added to ENZYME entries.

October 27, 1998

- The SWISS-PROT/TrEMBL and SWISS-2DPAGE full text search tools have been improved. The databases are now indexed using the Glimpse search engine, wildcards can be used in query strings, more fields (line types) are indexed and response times are much shorter than before.

- We have developed **NiceProt**, a tool that provides a user-friendly tabular view of SWISS-PROT entries (example). The 'NiceProt View of SWISS-PROT' is accessible from the bottom of each SWISS-PROT entry on ExPASy.

- The following database cross-references and literature references have been added to SWISS-PROT entries on ExPASy:
    - DR links to the PRESAGE resource for structural genomics from Stanford University (e.g. P53878);
    - DR links from relevant immunoglobin entries to IMGT, the international ImMunoGeneTics database from the University of Montpellier (e.g. P01876);
    - References to the Worm Breeder's Gazette in the RL lines of relevant entries from *Caenorhabditis elegans* (e.g. Q09517).

- Users who wish to save and retrieve all SWISS-PROT entries originating from a species can do this via the SWISS-PROT document 'List of organism identification codes': By clicking on any of the species codes (e.g. DROME) and specifying a filename, one can save all corresponding entries to a file which can be retrieved from the anonymous ExPASy FTP server.

- The output format of the PeptIdent peptide mass fingerprinting identification tool has been improved. PeptIdent results now contain a table summarizing information about the matching proteins, from where the user can jump to the detailed listing for the corresponding peptides.

- The new experimental tool CombSearch provides a unified interface for simultaneous queries to several protein identification programs accessible on the web. CombSearch was written by Rémi Hammerli and Pavel Dobrokhotov as a student project in the scope of the Bioinformatics course given at Geneva University.

- A new page providing links to conferences and events is available and accessible from the ExPASy home page. If you know about any conferences on molecular biology or bioinformatics we encourage you to register .

- The ExPASy interfaces which allow the direct submission of a SWISS-PROT/TrEMBL sequence to BLAST servers at EMBnet-CH and NCBI have been modified to provide a more transparent selection menu of BLAST programs and databases. These programs are designed for similarity searches easily accessible from a SWISS-PROT/TrEMBL entry; for advanced searches with more options we recommend to use the original BLAST submission forms at EMBnet-CH or NCBI.

August 24, 1998

- *There is a new tool in our section 'Protein identification and characterization tools':* PeptIdent allows the identification of proteins using pI, Mw and peptide mass fingerprinting data. Experimentally measured, user-specified peptide masses are compared with the theoretical peptides calculated for all proteins in SWISS-PROT. A species (or group of species) can also be specified for the search. PeptIdent makes extensive use of the annotations

in SWISS-PROT and takes into account post-translational modifications as documented in SWISS-PROT.
Results are displayed on-line or can be sent by email, in form of a html table. The result file contains direct links to FindMod to further characterize matching proteins by predicting potential protein post-translational modifications and finding potential single amino acid substitutions, and to PeptideMass.

- There is a new document describing how to create HTML links to services on ExPASy.
- In July 1998, SWISS-PROT, PROSITE and ENZYME have undergone major releases.
- New hypertext cross-references have been added to SWISS-PROT entries (example: P98073):
  - in RX lines: Medline abstracts corresponding to SWISS-PROT references can now also be consulted at the Weizmann Institute of Science in Israel, in addition to the archives at NCBI, ExPASy and GenomeNet Japan. These links have also been added to SWISS-2DPAGE entries.
  - DR DOMO lines have been added: These links provide direct access to relevant information in the DOMO database of homologous protein domains maintained by Jérôme Gracy at Infobiogen.
  - At the bottom of the page displaying a SWISS-PROT/TrEMBL entry, there are now direct links for submission of the sequence to ScanProsite and ProfileScan.
  - RL lines: Relevant SWISS-PROT entries are now directly linked to the Plant Gene Register, an electronic publication for articles describing the isolation and DNA sequence determination of plant genes (example: P48422).
  - The ExPASy interface to the BLAST server at EMBnet-CH now uses their new BLAST2 client, replacing WU-BLAST.

June 13, 1998

- The ExPASy server presents itself in a new layout: the home page, database entry pages, the tools page and many other pages have been redesigned for easier navigation and better readability.

  Users can now also use (in addition to the home page and ExPASy Index) the newly created clickable ExPASy site map to find useful tools, documents and services available on our server, and to find out about functional links between them.

  A new documentation page has been created which presents a complete table of documents available on ExPASy.

- There are two new SWISS-PROT documents:
  - humpvar.txt: an index of human proteins with sequence variants
  - humchr17.txt: an index of protein sequence entries encoded on human chromosome 17.
- Protein domains, chains etc. documented in the SWISS-PROT feature tables, if corresponding to subsequences of at least 10 amino acids, can now be directly submitted to a BLAST similarity search from the pages highlighting these subsequences. Example: DOMAIN EXTRACELLULAR ALPHA-1 (1A24 HUMAN).
- Two bugs have been corrected in ExPASy tools:
  - There was a small error in the computation of extinction coefficients by ProtParam: The contribution of Cysteines to the extinction coefficient (Gill S.C., von Hippel P.H. Anal. Biochem. 182:319-326(1989)) of a protein is only half of the values used previously in ProtParam, which results in slightly different values for the extinction coefficient.

- Our Translate tool no longer ignores base-ambiguity characters such as M, W, Y, etc. Previously performed translations for DNA sequences containing characters other than A,C,T,U,G, and N are likely to have been incorrect.

We apologize for any inconvenience caused by these errors and encourage our users to continue to send us their comments and bug reports.

March 27, 1998

- *There is a new tool in our section 'Protein identification and characterization tools':*
  FindMod is a program for the de novo discovery of protein post-translational modifications. It examines peptide mass fingerprinting results of known proteins for the presence of currently 18 types of PTMs of discrete mass. This is done by looking at mass differences between experimentally determined peptide masses and theoretical peptide masses calculated from a specified protein sequence. If a mass difference corresponds to a known PTM not already annotated in SWISS-PROT, "intelligent" rules are applied that examine the sequence of the peptide of interest and make predictions as to what amino acid in the peptide is likely to carry the modification.

- *Improved tools:*
  PeptideMass, which calculates masses of peptides and their posttranslational modifications for a given protein sequence, can now consider up to 3 missed cleavages. Post-translational modifications may be specified for a sequence in raw sequence format, and substitution tables are available to simplify the interpretation of the results for peptides concerned by database conflicts, variants or splicing variants.

  TagIdent can now search in SWISS-PROT, TrEMBL or both databases. It is also possible to perform an additional scan of a short sequence tag against all fragments contained in the database(s), even if pI and Mw cannot be computed for these proteins.

  MultiIdent (identification using pI, MW, amino acid composition, sequence tag and peptide mass fingerprinting data) is available for constellation 2 (Ala, Ile, Pro, Val, Arg, Leu, Ser, Asx, Lys, Thr, Glx, Gly, Met, His, Phe and Tyr. (Asp+Asn=Asx; Gln+Glu=Glx; Cys and Trp are not considered)) and constellation 4 (like constellation 2, but Gly is not considered).

- Several months ago, we started to distribute and update weekly, a set of data files that can be used to build a non-redundant protein sequence database consisting of SWISS-PROT, TrEMBL and TrEMBL updates. There is now a document explaining the contents and principles of this database.

- Information about the current release and update status of SWISS-PROT has been added to the SWISS-PROT page (currently 'Release 35 and updates up to 20-Mar-1998: 71198 entries').

- New hypertext cross-references have been added to SWISS-PROT entries:
  - in RX lines: Medline abstracts corresponding to SWISS-PROT references can now also be consulted on the Japanese GenomeNet server in addition to the archives at NCBI and ExPASy.
  - in DR PDB lines: Local copies of PDB entries are available. The user is now given the choice between accessing 3D structure information (e.g. 2hhe) in Geneva or Brookhaven (US). Both links provide direct access to 3D structure information in various formats, as well as hypertext links to servers offering related information.
  - DR PROTOMAP lines have been added: These links provide, for a SWISS-PROT entry, a cluster (group) of related proteins as classified by the ProtoMap server at Hebrew University, Jerusalem.

Example: DEFN_HUMAN.

- SWISS-2DPAGE is now available to be searched by the SRS Sequence Retrieval System.

February 13, 1998

- The SWISS-PROT full text search tool has been redesigned and improved. Boolean operators (AND, OR, NOT) can be used to combine and restrict queries, and special characters such as _ - # ' ( ) , . / are allowed as part of words (as used in SWISS-PROT).

- SWISS-PROT author names (RA lines) have been linked to a page listing all SWISS-PROT entries which contain references to articles (co-) authored by this author.

- The ExPASy interface to the EMBNet-CH BLAST server now contains a new option: This BLAST process manages two job queues: a (presumably) fast one and a slow one. Based on the sequence provided and the database requested, the process makes an (educated ???) guess to decide if the query will require more than 5 minutes of CPU time. Small jobs are allowed to proceed in the fast queue, while the others are forced to the slower one. If an e-mail address is provided, results of slow jobs will be automatically mailed back, while fast jobs will proceed as before.

- Two features have been added in SWISS-2DPAGE to facilitate visualisation and differentiation of spots:
  - If you click on a spot in one of the SWISS-2DPAGE maps (e.g. Plasma), the '2D' line describing this spot in the corresponding SWISS-2DPAGE entry is highlighted in green.
  - Hypertext links have been added from spot serial numbers on SWISS-2DPAGE '2D' lines to the master image for the protein, in which the spot with this serial number is highlighted in green (in contrast to the other spots displayed in red). Example: P00450.

January 13, 1998

- Since November 1997, SWISS-PROT, PROSITE, ENZYME and SWISS-2DPAGE have all gone through major releases.

- There is a new program that allows you to randomly retrieve a SWISS-PROT or TrEMBL entry.

- A new output format option has been added to our Translate tool. When translating a nucleotide sequence into a protein sequence, you can now also select to include, for each of the six open reading frames, the nucleotide sequence in the output.

- Cross-references and direct links to the Mendel Plant Gene Nomenclature Database have been added in corresponding SWISS-PROT entries. Example: P12084. There also is a file containing all SWISS-PROT entries with cross-references to Mendel in our series of "special selections", which is updated weekly and can be downloaded from our anonymous FTP server.

- Proteins which are documented to belong to an uncharacterized protein family in the SWISS-PROT CC (comments) lines, have been linked to the SWISS-PROT document upflist.txt. Example: P55061.

November 27, 1997

- In SRS (Sequence Retrieval System), SWISS-PROT DR (Database crossReference) and RC (Reference Comment) lines have been indexed. You may search for e.g. all entries with cross-references to PDB (enter 'PDB' in the DbName field), or all proteins that have been found in E.coli strain K12 (enter 'K12' in the 'RefComment' field).

- It is now possible to retrieve a number of SWISS-PROT/TrEMBL entries by specifying a list of accession numbers or entry names (ID).

- There are 4 new SWISS-PROT documents:
  - humchr18.txt: an index of protein sequence entries encoded on human chromosome 18;
  - pcc6803.txt: an index of Synechocystis strain PCC 6803 entries;
  - deleteac.txt: an index of deleted accession numbers.
  - upflist.txt: UPF (Uncharacterized Protein Families) list and index of members.

October 7, 1997

- We have implemented a search index, ExPASy Index, to help you find information within the ExPASy server. The index contains all the documents of ExPASy (currently about 800), except the database entries. It has been automatically indexed by the Marvin robot.
  Our new service BioHunt uses the same concept and allows you to search the internet for molecular biology information. In the current version, 17136 documents have been indexed.

- The ScanProsite tool has been modified to work with TrEMBL as well as SWISS-PROT. Furthermore, the part of the program which allows to scan a pattern against SWISS-PROT (and TrEMBL) has been improved and now avoids the previously frequent 'Document contains no data' error for large scan results.

- In PeptideMass, the set of post-translational modifications with discrete mass differences considered in peptide mass computation now also contains O-GlcNac (documented as FT CARBOHYD GLCNAC in SWISS-PROT) and C-Mannosylation of Tryptophan (FT CARBOHYD C-MANNOSYL). Thus, 17 post-translational modifications are now considered in PeptideMass. For examples, try CRAA_BOVIN or RNKD_HUMAN, don't forget to select "display all known post-translational modifications" and click on the "Perform" button.

- There is a new SWISS-PROT document:
  mgdtosp.txt - Index of MGD entries referenced in SWISS-PROT.

- Hyperlinks have been added from SWISS-PROT entries to the TIGR Microbial Database, which provides links to the information provided by TIGR on the genes encoded in the genomes they have sequenced (so far these are: Haemophilus influenzae, Helicobacter pylori, Methanococcus jannaschii, and Mycoplasma genitalium). (Example: FDHB_METJA)
  We have also created a specific file containing all SWISS-PROT entries containing cross-references to the TIGR database in our series of "special selections", which is updated weekly.

- SWISS-PROT reference (RL) lines and PROSITE references referring to one of the journals available at IDEAL, an online electronic library containing all 175 Academic Press journals, now contain direct links to the IDEAL server if the article was published in 1996 or later. From this, a 'Guest login' leads to the abstract of the article. (Example: RGSE_RAT)

September 5, 1997

Some new features of ExPASy:

- The PeptideMass program has been modified to take into account up to 2 missed cleavage sites. A new column 'MC' has been added to the output which indicates the number of missed cleavages, and peptides resulting from 0, 1 or 2 missed cleavages are displayed in different colours.
- A new parameter has been added in the ProtParam program: ProtParam results now include the grand average of hydropathicity (GRAVY) for a given protein.

- At the bottom of each SWISS-PROT and TrEMBL entry, there is now a link to a page displaying the entry in FASTA format (example: P11553).
- The local submission form to the WU-BLAST server at Lausanne has been changed to use as the default database the set of non-redundant protein databases SWISS-PROT, TrEMBL and TrEMBL_NEW.
- There are two new SWISS-PROT documents:
  - metallo.txt - Classification of metallothioneins and index of MT entries
  - hpylori.txt - Index of Helicobacter pylori strain 26695 chromosomal entries
- The display of current and previous Swiss-Flash bulletins has been redesigned: A table is available which lists all Swiss-Flash bulletins by category, including date, title and author of the bulletins.

July 24, 1997

We have now an SRS server (version 5) running on ExPASy. SRS (Sequence Retrieval System) allows you to retrieve entries across multiple databases with more sophisticated criteria than those allowed by the text-search interfaces available from the SWISS-PROT top page.

You can combine all the fields with logical operators and achieve queries like:

- Give me all vertebrate proteins having a PH domain and that are longer than 1000AA or
- Give me all calcium-binding proteins localized in the endoplasmic reticulum.

Five databases are indexed: SWISS-PROT, TrEMBL, TrEMBL_NEW, PROSITE, and ENZYME. SWISS-PROT and TrEMBL are updated on a weekly basis so that the set of these two databases stays non-redundant.

TrEMBL entries are now fully accessible on ExPASy via a cgi-script. The hypertext version of TrEMBL contains links to various databases and allows direct access to sequence analysis tools such as Swiss-Model, Blast, ProtParam, ProtScale, Compute pI/Mw and PeptideMass, as is the case for SWISS-PROT.

If you wish to link to a TrEMBL entry, you can use the following URL:

```
http://www.expasy.ch/cgi-bin/get-sprot-entry?<TrEMBL-AC>
```

e.g. to create a link to TrEMBL entry Q00061, use:

```
http://www.expasy.ch/cgi-bin/get-sprot-entry?Q00061
```

June 6, 1997

We are actively seeking any type of updates and/or corrections of SWISS-PROT entries, whether they have been published or not, and we encourage our users to submit us their suggested updates or corrections. This can be done using our new submission form, which can be accessed through an active link from the SWISS-PROT home page or from the bottom of each SWISS-PROT entry. Please read the tips and guidelines to find out what type of information we are seeking and how to proceed. We would already like to thank our users in advance for any contribution they can make in updating and correcting SWISS-PROT!

The tool which allows you to visualize and highlight the subsequence corresponding to a line in a SWISS-PROT feature table (FT) has been improved and is now using colour to highlight the subsequences in question. Example: in FA9_HUMAN:

```
FT   DOMAIN      93   129      EGF-LIKE 1, CALCIUM-BINDING (POTENTIAL).
```

May 21, 1997

At the bottom of each page displaying a SWISS-PROT entry, you will now find a link to a graphical Feature Table viewer (Java Applet) written by Thomas Junier at the Bioinformatics Group of ISREC Lausanne.

We have added several new hyperlinks in SWISS-PROT entries:

- The DR lines containing cross-references to EMBL/GenBank/DDBJ now include a link to a page displaying exclusively the corresponding CoDing Sequence (CDS).
- The RL lines referring to recent articles in certain journals whose WWW servers are maintained in collaboration with HighWire Press are now active hyperlinks to the abstracts of the corresponding articles. From the abstract page you can frequently access directly a full text on-line version of the article. The journals include J. Biol. Chem., Proc. Natl. Acad. Sci. USA, Science, Cell, etc.
- Entries with cross-references to MIM are now also linked (through a new virtual " DR GeneCards" line) to GeneCards, a database integrating information about the functions of human genes and their products, and of biomedical applications based on this knowledge. Example: BRC1_HUMAN.
- Entries belonging to family 1 of G-protein coupled receptors (as documented in feature tables) now contain active links to GPCRDB-Snakes diagrams (through the new virtual " DR GPCRDB-Snakes" line) prepared by the GPCRDB group at EMBL Heidelberg. Example: 5H1A_HUMAN.

There are 3 new SWISS-PROT documents:

- humchr19.txt: an index of protein sequence entries encoded on human chromosome 19
- ngr234.txt: a table of putative genes in Rhizobium plasmid pNGR234a
- initfact.txt: a list of translation initiation factors

On the ExPASy anonymous FTP server, the SWISS-PROT update files new_seq.dat, upd_ann.dat and upd_seq.dat are now also available in compressed form in the directory /ftp/databases/swiss-prot/updates_compressed/.

March 27, 1997

We have modified and improved access from ExPASy to various BLAST (Basic Local Alignment Search Tool) similarity search services:

In the tools page, you can now choose between 5 different interfaces to BLAST servers in Switzerland, the USA and Germany:

Switzerland:

Running on a 2-processor Pentium Pro machine, the new WU-BLAST server at EMBNet Switzerland in Lausanne has a faster response time than the EPFL server, and should be more stable. As opposed to the original NCBI BLAST algorithm, WU-BLAST generates gapped alignments. A full set of weekly updated databases is provided.

- Local interface to WU-BLAST at EMBNet-CH (Lausanne)
- Original interface to WU-BLAST at EMBNet-CH (Lausanne)

USA:

- Local interface to BLAST at NCBI
- Original interface of BLAST at NCBI

Germany:

- WU-BLAST at Bork's group in EMBL (Heidelberg)

For direct BLAST submission from a SWISS-PROT entry (icons at the bottom of the page displaying an entry - example), you have the choice between the servers at NCBI and EMBNet-CH.

The following documents have been added to the list of SWISS-PROT documents:

- bloodgrp.txt - Blood group antigen proteins
- fly.txt - Index of Drosophila entries and their corresponding FlyBase cross-references
- mjannasc.txt - Index of Methanococcus jannaschii entries
- mgenital.txt - Index of Mycoplasma genitalium strain G-37 chromosomal entries

March 17, 1997

We have completely rewritten the Swiss-Shop sequence alerting system for SWISS-PROT that allows you to automatically obtain (by email) new sequence entries relevant to your field(s) of interest.

In the new version of Swiss-Shop, some new features have been added:

As before, you can either launch a sequence/pattern based search or a keyword based search.

- For a sequence based search, you need to specify a SWISS-PROT ID or AC or a raw protein sequence, and your sequence will be scanned, at each weekly update of SWISS-PROT, against the new sequences in the database using the alignment program BLAST. Sequences thus found to be similar to your protein will be sent to you by email. It is up to you to specify the BLAST probability threshold for P(N) (the probability that the alignment is real and not random), and you will receive a list of all sequences for which this probability is below the specified value.
- For a pattern based search, enter a PROSITE ID or AC or a pattern in PROSITE format, and Swiss-Shop will scan this pattern, at each weekly update of SWISS-PROT, against the sequences that have been added in SWISS-PROT since the last weekly update. You will receive the list of new entries matching your pattern.
- For a keyword based search, it was previously possible to specify keywords from SWISS-PROT OS, OC, OG (taxonomy), RA (authors), KW, DE, CC lines. In addition to these lines, you can now also search DR (Cross-references to other databases) and FT (feature) lines with one or more specified keywords. Swiss-Shop will look for these keywords on the corresponding lines of all SWISS-PROT entries added in the database since the last weekly release.

Furthermore, we now offer you 4 different output formats. You can choose to receive the sequences matching your query

- as a file in SWISS-PROT format or
- as a list of SWISS-PROT accession numbers or
- in form of a short report containing information from SWISS-PROT ID, AC, DE, OS lines or
- as a list of SWISS-PROT accession numbers with hypertext links to the corresponding entries on the ExPASy WWW server. This allows you to view your email message with your Web browser and to follow the hypertext links to the full entries on ExPASy.

You can further specify if you wish to be notified every time Swiss-Shop is run, even if there are no new sequences matching your query, or to receive an email report only when there are new SWISS-PROT entries matching your search terms.
You can specify the expiration date of your request, the default being one year after submission.
For editing previous requests (e.g. to update the expiration date or to modify search criteria) you can enter a password for each new request. This allows you to open the request later and edit it on-line rather than deleting it and submitting a new one.

March 6, 1997

New and improved protein identification tools:

There is a new tool on ExPASy:

- MultiIdent: This tool achieves protein identification using parameters such as protein species, estimated pI and MW, AA composition, sequence tag, and peptide mass fingerprinting data. It is particularly suited to the identification of proteins across species boundaries. Currently, the program works by first generating a set of proteins in the database with AA compositions close to the unknown protein, as for AACompIdent. Theoretical peptide masses from the proteins in this set are then matched with the peptide masses of the unknown protein to find the number of peptides in common (number of "hits"). Three types of lists are produced in the results. Firstly, a list where proteins from the database are ranked according to their AA composition score; secondly, a list where proteins are ranked according to the number of peptide hits they showed with the unknown protein; and thirdly, a list that shows only proteins that were present in both the above lists, where these proteins are ranked according to an integrated AA and peptide hit score. In all these lists, protein pI, MW, and species of origin (using a term from SWISS-PROT OS or OC lines) and keywords can be used, as in AACompIdent, to increase the specificity of searches.

The following tools have been improved, offering numerous additional features:

- AACompIdent (identification of a protein from its amino acid composition)
  You can restrict your search by specifying one or more term(s) from the OS or OC lines of SWISS-PROT (example: *HOMO SAPIENS* or *MAMMALIA*). You can also enter a keyword appearing on the KW lines of SWISS-PROT to further restrict your search. For example, a keyword of "CALCIUM-BINDING" could be used in conjunction with the OC term "MAMMALIA" to see if a user- entered protein matches well with any mammalian calcium-binding proteins in the database.
- TagIdent now allows, for one or more species (term from SWISS-PROT OS or OC lines) and with an optional keyword,
  1. the generation of a list of proteins close to a given pI and Mw,
  2. the identification of proteins by matching a short sequence tag of up to 6 amino acids against proteins in the SWISS-PROT database close to a given pI and Mw,
  3. the identification of proteins by their mass, if this mass has been determined by mass spectrometric techniques.

For PeptideMass, Compute pI/Mw, AACompSim and all the above-mentioned tools, documentation and references have been added and the submission forms have been reformatted and improved.

March 4, 1997

Thanks to the generosity of the Geneva Government, we have been able to acquire a new computer for the ExPASy server (a Sun Microsystems Ultra Server Enterprise 2). The server is now accessible at URL:

http://www.expasy.ch

The old URL remains valid for some time.

January 9, 1997

Some new features of ExPASy:

- New active links have been established from SWISS-PROT entries
  - to the TRANSFAC database of transcription factors;
  - from Bacillus subtilis entries to Micado (MICrobial Advanced Database Organization) at

- INRA, France;
    - to local copies of MEDLINE abstracts. We now give the user the choice of retrieying a MEDLINE abstract (example: 90368558) from either NCBI or Geneva;
    - to our Peptide Mass tool which cuts a protein sequence with a chosen enzyme and computes the masses of the received peptides.
- From Release 35 on, SWISS-PROT comments (CC) lines can contain a new 'topic' "DATABASE", which contains information about related databases catering for a specific protein or a for a very limited number of proteins. Most of these databases are mutation databases, reporting defects linked to a genetic disease. If such a database is available electronically, the CC DATABASE lines provide the relevant electronic coordinates, e.g. in P29965 (CD4L_HUMAN):

```
CC   -!- DATABASE: NAME=CD40Lbase; NOTE=European CD40L defect database;
CC        WWW="HTTP://www.expasy.ch/cd40lbase/";
CC        FTP="ftp.expasy.ch/databases/cd40lbase".
```

- There is a new SWISS-PROT document:
  yeast13.txt - a list of Yeast Chromosome XIII entries.
- Two new features have been added in ENZYME entries:
    - direct links from an enzyme to all relevant maps of Boehringer Mannheim's Biochemical Pathways and
    - links to the WIT (What Is There) database of metabolic pathways.

November 26, 1996

The Boehringer Mannheim Biochemical Pathways maps and index have been digitised and are now accesible on this server. Enter a keyword (such as, for example *Oxoacyl*) and surf on the biochemical pathways maps.

November 11, 1996

CD40Lbase, The European CD40L Defect Database prepared by Manuel Peitsch, has been made accessible through this server. The purpose of CD40Lbase is to collect clinical and molecular data on CD40 ligand defects leading to X-linked Hyper-IgM syndrome.

A new tool is available from the Tools page: The PeptideMass Peptide Characterisation Software. This program is designed to calculate the theoretical masses of peptides generated by the chemical or enzymatic cleavage of proteins, to assist in the interpretation of peptide mass fingerprinting and peptide mapping experiments. Protein sequences can be provided by the user or can be a code name for a protein in the SWISS-PROT protein database. When proteins of interest are specified from SWISS-PROT, the program considers all annotations for that protein in the database, and uses these in order to generate the correct peptide masses and warn users about peptides that are not likely to be found when undertaking peptide mass fingerprinting. Many protein post-translational modifications which affect the masses of peptides can thus be taken into consideration.

In PROSITE and Enzyme, we have added the possibility to save all referenced SWISS-PROT entries to a file on our anonymous FTP server (in the outgoing directory).

The Compute pI/Mw tool has been included in the list of sequence analysis tools that can be directly accessed from a SWISS-PROT entry.

Two new SWISS-PROT documents are available:
- humchr20.txt - an index of protein sequence entries encoded on human chromosome 20
- tisslist.txt - a list of the currently valid values for the "TISSUE" topic of the RC line type in SWISS-PROT.

September 30, 1996

A new SWISS-PROT document has been added: ribosomp.txt - an index of ribosomal protèins classified by families on the basis of sequence similarities.

In ec2dtosp.txt, an index of E. coli Gene-protein database (ECO2DBASE) entries referenced in SWISS-PROT, we have established direct links to ECO2DBASE, and SWISS-PROT entries now also contain links to ECO2DBASE.

At the end of each page displaying a SWISS-PROT entry we have added links to our sequence analysis tools ProtParam and ProtScale,which allows the user to directly submit the SWISS-PROT sequence to these tools.

September 19, 1996

Some new features of ExPASy:

- We have created a new protein identification tool called TagIdent. This is a modification of the old tool **GuessProt**. The user can now identify proteins from 2-D gels by giving protein pI and MW estimates, a species or organism classification of interest, and a short sequence tag of up to 6 amino acids. This tag can be derived from the N-terminus, the C-terminus or from internal peptides of a protein. The results are now sent to the user by e-mail, allowing many searches to be done at the same time. If you only want to generate a list of potential proteins in a specific pI or MW range (as was the function of the old tool GuessProt), do not select the TAG option in the form.
- An email option has been added to the tool ScanProsite: if you want to scan a pattern against SWISS-PROT, you have now the option of having sent the results of your query by email, which should avoid previously frequent timeout problems and is particularly useful for complex patterns.
  ScanProsite, which only scans SWISS-PROT with PROSITE *pattern* entries (as opposed to *rule* and *matrix* entries), can now also be used with the PROSITE rule entry PS00013, PROKAR_LIPOPROTEIN.
- SWISS-PROT entries have been linked to DDBJ, the DNA Data Bank of Japan. We have also added direct links to the Bacillus subtilis genomic data bank, SubtiList and to the Yeast Protein Database YPD to relevant SWISS-PROT entries.
- Links have been established from most feature (FT) lines of SWISS-PROT entries to pages that highlight the subsequence in question, both in 1- and in 3-letter amino acid codes.
  Example: in FA9_HUMAN:

```
FT   DOMAIN      93    129       EGF-LIKE 1, CALCIUM-BINDING (POTENTIAL).
```

- We have added three new SWISS-PROT documents:
  humchrx.txt - an index of protein sequence entries encoded on human chromosome X
  yeast7.txt - a list of Yeast Chromosome VII entries
  yeast14.txt - a list of Yeast Chromosome XIV entries.
- 2D Hunt, a database created and continuously updated by the Marvin robot contains sites related to electrophoresis and specifically to 2-D electrophoresis. It is now searchable from the SWISS-2DPAGE top page.

April 11, 1996

AACompIdent: New options - AACompIdent is a tool which allows the identification of a protein from its amino acid composition. It searches SWISS-PROT for proteins, whose amino acid compositions are closest to the amino acid composition given. Two new options and a new constellation have been added to this tool:

A. C-Terminal display in tagging option

The user may now choose between displaying the C or N terminal side of the proteins that score best.

B. Permutation search in tagging option

This option searches for all permutations of the given tag in the sequences.

C. Constellation 4

Constellation 4 has been added: Ala, Ile, Pro, Val, Arg, Leu, Ser, Asx, Lys, Thr, Glx, Met, His, Phe and Tyr. (Asp+Asn=Asx; Gln+Glu=Glx; Gly, Cys and Trp are not considered).

March 22, 1996

We have added a new tool, ProtScale which allows you to compute and represent the profile produced by an amino acid scale on a selected protein. 50 scales are provided, including 'classics' such as the Kyte and Doolittle hydrophobicity scale.

Links have been added between relevant SWISS-PROT entries and the 2D gel protein databases at Harefield.

A new SWISS-PROT document has been added which describes the nomenclature of glycosyl hydrolases (GH) and that includes an index of sequences that belong to the various GH families.

A PC (MS-Windows) version of LALNVIEW (graphical viewer for pairwise alignments) is now available.

Nicolas Guex has produced a new logo for PROSITE.

February 16, 1996

We have added a new tool, SIM which computes a user defined number of best non-intersecting alignments between two sequences. The results of the alignment can be viewed graphically using the LALNVIEW program developed by Laurent Duret and which is currently available for Macs and UNIX.

Additional links have been added in the tools page, notably to the Weizmann Institute ultra-fast rigorous (Smith/Waterman) similarity searches using the Bioccelerator and to the Garnier, Osgoodthorpe and Robson (GOR) secondary structure prediction method at SBDS.

The SeqAnalRef database now includes a section listing author's email and eventually also WWW home pages. It is also possible to access the links from a page displaying either a reference list or a single reference.

Amos has recently started to create a list of Biomolecular servers for his own usage, but as some people have asked to access this list (which is under construction), we are making it available from the ExPASy top page. Many other small changes were carried out in the last two months.

We thank you for using ExPASy (we have now reached a cumulative total of 4 million connections).

December 14, 1995

After 29 months of existence the ExPASy molecular biology server received a new logo, designed and produced by Nicolas Guex.

October 23, 1995

The Melanie page has been reorganised. With the announcement of release 2.1 of the **Melanie II 2-D PAGE analysis software package**, a complete up-to-date description of the software as well as a comprehensive tutorial are now available.

October 13, 1995

Links have been added between SWISS-PROT Escherichia coli K12 chromosomal entries and the EcoCyc database, the encyclopedia of E. coli Gene and Metabolism.

You can now seach in PROSITE by citation.

October 9, 1995

Some new features of ExPASy:

- Search in SWISS-PROT by citation - When you call this option, you are prompted to enter the name of a journal and optionally a volume number and/or a year. The program is written in such a way that you can enter either the full name of a journal or its official abbreviation.
- RandSeq - a new tool to generate random protein sequences.
- SWISS-PROT document haeinflu.txt - Index of Haemophilus influenzae RD chromosomal entries and gene names with links to the TIGR and EMBL servers.
- SWISS-PROT document submit.txt - Description of how to submit sequence data to the SWISS-PROT data bank.
- SWISS-PROT document aatrnasy.txt - List of aminoacyl tRNA synthetases.
- Swiss-Jokes - A new page to give access to our collection of jokes from the fields of molecular biology and of computing.

Many other changes have been done, such as the redesign of the Geneva local pages; the addition, in the tool page, of a link to ProfileScan.

It should also be noted that when you search in SWISS-PROT by either description or by full text and that your seach criteria returns more than two entries, you can save these entries to a file on our anonymous FTP server (in the outgoing directory).

September 19, 1995

AACompIdent: New options - AACompIdent is a tool which allows the identification of a protein from its amino acid composition. It searches SWISS-PROT for proteins, whose amino acid compositions are closest to the amino acid composition given. A new option and a new constellation have been added to this tool:

A. Tagging option

With this option, the first 40 amino acid of each protein are printed in the result, instead of the protein name. One may optionnally also enter a tag (a short seuqnece, typically 3 to 8 residues) which will be matched with the sequences of the resulting proteins. Proteins matching the tag will be marked.

B. Free constellation

This is a free constellation, that is one may select any amino acid constellation he/she likes. One just have to fill in the composition values for the selected amino acids. The values will then be normalised, so that the total make 100 (percent).

September 4, 1995

A new page has been created: WORLD-2DPAGE is an index to all known federated 2-D PAGE database servers, as well as to 2-D PAGE related servers and services.

July 22, 1995

A new tool has been implemented on ExPASy, ProtParam allows the computation of various physical and chemical parameters for a given protein stored in SWISS-PROT or for a user entered sequence. The computed parameters include the molecular weight, theoretical pI, amino acid composition, extinction coefficient, estimated half-life, instability index and aliphatic index

The Journal of Biological Chemistry (JBC) has a WWW server where abstracts and full text of articles are made available. We are happy to announce the implementation of what we believe to be the first direct link in a sequence database between a reference and the full text version of a cited article. Recent JBC references are directly linked to the corresponding entry point in the JBC server. If you want to see such a link, take a look at reference 3 in SWISS-PROT entry KDSA_ECOLI.

The SWISS-PROT document file jourlist.txt which provides information on all the journals cited in that database, now contains links to WWW or Gopher servers set up by a variety of publishers of academic journals.

Two new SWISS-PROT document have been added, one is a nomenclature and index of peptidase sequences, the other is the list of Yeast Chromosome VI entries in SWISS-PROT

June 19, 1995

A new tool has been implemented on ExPASy, ScanProsite allows to either scan a protein sequence the occurence of patterns stored in the PROSITE database or to scan the SWISS-PROT database - including weekly releases - for the occurence of a pattern.

We are happy to announce a new ""service"" Swiss-Quiz The principle of this quiz is to answer to 10 randomly chosen questions relative to the fields of molecular biology, biochemistry and genetics. Each month, we will randomly pick up one person among all those that have obtained a perfect score (and it's not so easy !) and will send that person some delicious Swiss chocolate !

Links have been added from SWISS-PROT to the Saccharomyces genomic database (SacchDb) at

Stanford.

A new SWISS-PROT document has been added, it is a nomenclature and index of allergen sequences.

May 26, 1995

A new service is available: SWISS-2DSERVICE. The Two-Dimensional Gel Electrophoresis Laboratory of Geneva, Switzerland, is running a highly reproducible method for the two-dimensional separation of proteins. The laboratory now provides a 2-D PAGE service to which you may send your samples for analysis. This service includes analytical and preparative high-resolution 2-D PAGE, electrotransfer on membranes and/or amino acid composition.

May 17, 1995

New link in the Tools page to the multiple sequence alignment at Washington University.

May 11, 1995

Two links have been added to the SWISS-PROT entries. The first one directly submits a request to Swiss-Model for a 3D model of the current SWISS-PROT protein. The result is then sent back by e-mail. The second one allows to perform a sequence alignment with the current sequence, using NCBI's Basic Local Aligmment Search Tool. This link is especially interesting in the virtual SWISS-PROT entries produced by the Translate tool.

May 5, 1995

We announce a new service, SWISS-FLASH, that reports news of databases, software and services developments from the Swiss biocomputing groups responsible for the ECD, ENZYME, LISTA, PROSITE, SeqAnalRef, SWISS-2DPAGE, SWISS-3DIMAGE and SWISS-PROT databases; the Melanie software package; the WWW ExPASy server; the SWISS-Model, SWISS- Shop and other network-based computational tools; and the SWISS-2DSERVICE services. If you subscribe to this service, you will automatically get the SWISS-Flash bulletins by electronic mail.

The SWISS-3DIMAGE database has been completely reorganised and indexed. The database is now searchable in the same way as the other SWISS-*** databases. We now also supply pictures in JPEG format, in addition to GIF and SGI. The images may still be downloaded by FTP.

Links to REBASE points now the version maintained at John Hopkins, whose layout is nicer than our own text based version !

April 19, 1995

We added Translate, a new tool which allows the translation of a nucleotide (DNA/RNA) sequence to a protein sequence.

Most of the pages in the server have been "refreshed" to make them more readable.

March 21, 1995

Links have been added from SWISS-PROT to the LISTA database of budding yeast (Saccharomyces cerevisiae) genes coding for proteins prepared under the supervisation of Patrick Linder.

March 7, 1995

Links have been added from SWISS-PROT to the HSSP database of structure-sequence alignments from the Protein Design Group, EMBL, Heidelberg.

March 2, 1995

During the last two months, various links have been added:

- from SWISS-PROT to the SubtiList and YEPD databases
- from ENZYME to PROSITE and to the Ligand database in Kyoto
- internally from PROSITE entries to other relevant PROSITE entries

Links from SWISS-PROT to FlyBase use the new WWW server for that database.

Many new SWISS-PROT documents have been added.

The page on the Melanie 2-D PAGE analysis software has been completely redesigned and includes now a on-line tutorial, as well as a request for information form.

December 7, 1994

In order to help users navigate through the BxPASy server, we have added graphical examples. More will be added in the future. See for example: Celegans examples or the who's who on ExPASy page. Thanks to Brigitte Boeckmann for the illustrations.

October 31, 1994

ENZYME: the *ENZYME Data Bank* has been added to the ExPASy server, This database may be accessed by EC number, name, compound, cofactor, comment, or by browsing through the list of classes, subclasses and sub-subclasses. Any entry in SWISS-PROT that contains an EC number in the DE line has also a direct link to ENZYME (by clicking on the EC number).

October 20, 1994

New services:

- Swiss-Shop - a sequence allerting system for Swiss-Prot that allows you to automatically obtain new sequence entries relevant to your field(s) of interest.
- Swiss-Model - an automated knowledge-based protein modelling server.

Compute pI/Mw: the tool to compute pI and Mw now accepts also a list of ID/AC's.

SWISS-PROT: in PDB cross-reference lines, there is now a link called RASMOL, sending the PDB entry as a *chemical/pdb* MIME type. On Unix systems, if you add, in the file .mailcap in your home directory, a line of the form

```
chemical/pdb; rasmol %s
```

then RASMOL will automatically be launched to display the protein 3D structure. This works also with any other program which accepts PDB coordinates. On systems other than Unix, this may also be specified. See your browser's manual.

October 13, 1994

The SWISS-PROT top page has been re-modeled. A number of new functionalities and documents have been added.

October 7, 1994

New tools have been added:

- **Amino acid composition similarity search** - the search may now also be performed from a given SWISS-PROT entry, whose amino acid composition will be compared with the whole SWISS-PROT database.
- **Compute pI/Mw** - Compute the theoretical pI and Mw from a SWISS-PROT ID or AC, or for a given sequence.

October 5, 1994

The gels run during the 2-D PAGE courses in Geneva are now displayed on the server.

September 29, 1994

SWISS-2DPAGE: protein maps now have a pI/Mw scale.

SeqAnalRef: the *Sequence Analysis Bibliographic Reference* database has been added to the ExPASy server. This database may be accessed by keyword, by reference identifier, by author and by full text search.

List of on-line experts: in SWISS-PROT and PROSITE top pages, a list of on-line experts gives you the possibility to directly send questions to any of the listed experts. The list is ogranized by subjects.

SWISS-PROT: new lists added:

- List of abbreviations for journals cited
- List of species has been made active
- Yeast Chromosome III entries in SWISS-PROT
- Nomenclature of extracellular domain
- List of on-line experts

PROSITE: new 3D line with active links to PDB.

September 26, 1994

In the tool AACompIdent for identifying a protein by its amino acid composition, options have been added. They allow to specify how many proteins should be displayed, as well as the pI and Mw range in which the search should be performed.

Also, some old bugs have now been corrected.

September 12, 1994

The tool AACompIdent for identifying a protein by its amino acid composition, has been corrected and is now supposed to work. If you still encounter problems, please send us a mail.

June 17, 1994

SWISS-PROT: added cross-references (DR lines) to **GenBank**.

June 16, 1994

SWISS-PROT: added cross-references (DR lines) to **MaizeDB** Maize Genome Database of the National Agricultural Library.

June 6, 1994

Added the PROSITE page: PROSITE entries may now be searched by description of sites and pattern, by accession number, by author, and soon by full text search.

June 3, 1994

Added the GuessProt tool to the tools page: you may now get the SWISS-PROT proteins closest to a given *pI* and *Mw*.

May 27, 1994

In SWISS-PROT entries, added links to **GCRDb** - the *G-Protein--Coupled Receptor DataBase*.

Added the **list of nomenclature related references for proteins** to the SWISS-PROT top page.

May 26, 1994

Added a new reference 2-D PAGE map of **Platelet** to SWISS-2DPAGE.

May 20, 1994
The SWISS-2DPAGE team is now organizing a **2-D PAGE training** in Geneva once every three months.

May 18, 1994
Added the **Yeast Chromosome XI** list of proteins to the SWISS-PROT documentation page.

May 11, 1994
Tools: new page giving access to on-line analysis tools, such as BLAST, BLITZ, PROSITE search and amino acid composition analysis, and more to come in the future.

March 23, 1994
Added the list of **restriction enzymes and methylases** in SWISS-PROT top page.

March 22, 1994
The ExPASy WWW server has been upgraded to a **SPARCServer 10/51**. It should perform much faster now. If some features are not working, please tell us about.

March 18, 1994
The links to OMIM are now direct links to the **OMIM hypertext** server from GDB. Thanks to Keith Robison for informing me about it.

March 4, 1994
SWISS-2DPAGE: Added experimental **Amino Acid Composition Similarity Search** : you enter a protein's amino acid composition and the server will e-mail you the list of SWISS-PROT entries with similar compositions, sorted by decreasing similarity measure.

March 2, 1994
Added direct link to NCBI's **BLAST** Basic Local Alignment Search Tool (ExPASy and SWISS-PROT top pages).

March 1, 1994
Starting with release 28, **SWISS-PROT keyword search** will be performed on the main release as well as on the **weekly updates**.

In the SWISS-PROT page, added links to four additional active lists:

- Index of Escherichia coli K12 chromosomal entries in SWISS-PROT and their corresponding EcoGene cross-reference
- Index of Saccharomyces cerevisiae entries in SWISS-PROT and their corresponding gene designations
- Index of Caenorhabditis elegans entries in SWISS-PROT and their corresponding gene designations and WormPep cross-references
- Index of Dictyostelium discoideum entries in SWISS-PROT and their corresponding gene designations and DictyDB cross-references

February 23, 1994
Added two new reference 2-D PAGE maps: **Macrophage Like Cell Line** (U937) and **Erythroleukemia Cell** (ELC).

In a SWISS-2DPAGE entry, it is now possible to compute the **theoretical pI and Mw** of the protein.

February 14, 1994

Added **SWISS-2DPAGE Map Selection** : you select a 2-D PAGE reference gel, click on a spot and get information on the corresponding protein. See the SWISS-2DPAGE top page.

February 11, 1994

Added a new reference 2-D PAGE map of **Cerebrospinal Fluid** to SWISS-2DPAGE.

January 28, 1994

Added the **bionet newsgroups**.

January 25, 1994

Added an entry to **SWISS-3DIMAGE** images of crystallized proteins.

In SWISS-PROT entries which contain cross-references to PDB, added a cross-reference to **SWISS-3DIMAGE**. Try for example *AAT_ECOLI*.

January 24, 1994

Added **full text search** of the SWISS-PROT protein sequence database.

January 17, 1994

Added links to **MEDLINE** entries in SWISS-PROT, through NCBI's Entrez Server.

Added, in the SWISS-2DPAGE page, a link to the **QUEST Protein Database Center**.

December 1, 1993

Added a **User Survey**. Please help us inprove the server in participating to this survey.

Added a new reference 2-D PAGE map of **Lymphoma** to SWISS-2DPAGE.

November 23, 1993

Added link to **BioBit 24, the BIO-NAUT Newsletter** from November, 22, 1993, describing the World Wide Web.

November 18, 1993

Added links to the **Maize Genome Database** at Columbia, Missouri and to **EMBnet Switzerland**.

November 17, 1993

Added the list of overall **Top Ten** users in the *ExPASy server Activity Reports* page.

November 16, 1993

Added **Images of crystallized proteins** from this server.

Added links to **Harvard Biological Laboratories**, the **Gene-Server** at University of Houston, the **EMBnet**: Biocomputing in Europe, the **biology servers index** at USGS, **Jackson Laboratory** WWW server and Keith Robison's **Molecular Biology WWW sampler**.

October 12, 1993

Added a list of **specialised documents** to the SWISS-PROT top page, such as 7-transmembrane G-linked receptors, CD nomenclature for surface proteins of Human leucocytes and Vertebrate homeobox proteins. Some of these list give then direct access to corresponding SWISS-PROT entries.

October 8, 1993

Added links to the **Caenorhabditis elegans** and **Mycobacterium** databases at INRA (France).

Added a link to the ExPASy **server activity** reports.

October 4, 1993
Moved to the NCSA server.

September 28, 1993
Added the **PDB** Brookhaven Protein Data Bank of 3D structures. In SWISS-PROT, cross-references to PDB have now active links to the gopher server at Protein Data Bank. You may access the PDB entry or get the 3D image. Try for example the SWISS-PROT entry *P00782*.

September 27, 1993
Added the **FlyBase** database of genetic and molecular data for Drosophila. In SWISS-PROT and EMBL, cross-references to FLYBASE are now active links. Therefore, SWISS-PROT has now active links to SWISS-2DPAGE, EMBL, PROSITE, REBASE, OMIM and FLYBASE. EMBL has active links to SWISS-PROT and FlyBase.

September 23, 1993
Added a link to the National Institute of Health **Genobase** server to our top page.

September 21, 1993
Announced the ExPASy server and SWISS-2DPAGE release 0 to bionet.announce.

August 1, 1993
Installed the ExPASy molecular biology server, release 0, beta version.

Last modified 21/Oct/2004 by CHH

ExPASy Home page **Site Map** **Search ExPASy** **Contact us**

Hosted by CBR Canada Mirror sites: Australia Brazil Korea Switzerland Taiwan USA

What's New Archive

PubMed Entrez BLAST OMIM Books TaxBrowser Structure

NCBI

SITE MAP

What's New Archive

2001
2000
1999
1998
1997
1996
1995
1994

BACK

1996

11/25	Electronic PCR	Electronic PCR is now available. PCR-based sequence tagged sites (STSs) have been used as landmarks for construction of various types of genomic maps. Using "electronic PCR" (e-PCR), these sites can be detected in DNA sequences, potentially allowing their map locations to be determined.
11/14	Sequin, Release 1.71	A new release of Sequin, a sequence submission tool, is now available. Version 1.71 features improved handling of phylogenetic sets of sequences and also allows users to update their own pre-existing database records.
11/04	dbGSS Announced	The Database of Genome Survey Sequences (dbGSS) is now available. This database contains more detailed information than the corresponding records in the GSS Division of GenBank.
10/24	Human Gene Map	The Gene Map of the Human Genome published in the October 25 issue of *Science* is available. This map shows the chromosome location of over 16,000 human genes with links to the underlying sequence and map data.
10/04	Sequin	Sequin, a stand-alone sequence submission tool, has a new release with several enhancements, including a repeat finder and ORF finder. New documentation and a tutorial are available, both on the Web and in NCBI's newsletter.
09/27	ORF Finder	The Open Reading Frame (ORF) Finder is a graphical analysis tool that finds all open reading frames in a user's sequence or one already in the database of a selectable minimum size.
09/06	Virological Software	Software for analyzing animal trials and calculating infectious and 50% inhibitory doses is now available. The programs VacMan and

		ID-50 can now be downloaded as self-extracting archives for either IBM or Macintosh computers.
08/23	Complete Genome, *Methanococcus jannaschii*	The complete genome sequence and annotation of *Methanococcus jannaschii*, prepared by The Institute for Genomic Research (TIGR) is now available in Entrez Genome, as well as in GenBank, where the 1.7-megabase sequence has been separated into 150 records of approximately 11,000 bp each. The graphical view (as well as a link to underlying data) of the complete genome is present in Entrez Genome, along with the extrachromosomal elements 1 and 2. The complete sequence is also available by anonymous FTP; see the README file for a description of the various files in the genomes division directory.
08/20	Batch Entrez	Downloading large numbers of sequence records from Entrez is now possible through 'Batch Entrez'. User can specify a download for an entire set of records for a given organism or for a set of accession numbers. The data are saved to a file on the user's computer.
08/05	*Saccharomyces cerevisiae* Database	A new database has been added to the BLAST databases: all the nucleotide sequences from the yeast (*Saccharomyces cerevisiae*) genome sequencing project and their encoding amino acid sequences can now be searched with the BLAST suite of programs.
07/26	Cn3D in Entrez	A major new release of Network Entrez is now available. Release 5.0 contains Cn3D, a new 3D structure viewer integrated into Network Entrez.
07/15	BLAST2	The BLAST2 network service is now available on the FTP site without registration. Three clients for multiple platforms are available: blastcli has a convenient graphical interface and produces the "traditional" BLAST output; blastcl2 is a command-line client (meant mostly for UNIX) that also produces the traditional BLAST output; and PowerBlast produces a one-to-many alignment, allows filtering by organism, and allows a gapped alignment as a post-processing of the BLAST results. Users of the older Experimental BLAST Network Service (with the exception of GCG users, who are still required to register

		and use the older program) are encouraged to switch to this newest version.
05/21; see also 11/14	Sequin	Sequin is a program for submitting and updating GenBank entries. It is designed to simplify the sequence submission process, provide graphical viewing and editing options, and allow submission of segmented entries. Sequin automatically adjusts feature table positions as the sequence is edited. Versions of Sequin are available through FTP for the Macintosh, PC/Windows, UNIX, and VMS.
05/21	PowerBlast	PowerBlast is a new network BLAST application for automated analysis of genomic sequences. It combines BLAST searching with filtering for low complexity regions and repeats. It can generate organism-specific output and compute optimal, gapped alignments. The results are displayed graphically and textually as multiple alignments, with annotated features superimposed on the aligned sequences. Versions of PowerBlast are available through FTP for the Macintosh, PC, SunOS, and Solaris.
05/06	WWW BLAST	The WWW BLAST page has been extensively revised. It now has both a simplified "Basic" Blast Search, allowing a user to search with the default parameters, as well as an "Advanced" page, where users may set BLAST parameters. An email option allows a user to receive results in a convenient form.
04/10	WWW Entrez	WWW Entrez now provides graphical views of nucleotide and protein sequences and access to the NCBI Genomes database, which contains graphical views of sequences and chromosome maps. Click on "Graphical view" from an Entrez document summary or click on the "Graphic" button from a sequence report.
03/12	Mouse/Human Homology	The Seldin/Debry Mouse/Human Homology Relationships page presents a table comparing genes in homologous segments of DNA from human and mouse sources, sorted by position in each genome.
03/07	Complete Genomes	An NCBI research project, Complete Genomes, presents the results of analyses of complete genome sequences. The analyses for the genomes of *Haemophilus influenzae*, *E. coli* (75%), and *Mycoplasma genitalium* are

		now available.
03/08	BLAST Databases	Changes to the BLAST Databases (February 20 announcement superseded by that of March 8.)
02/15	Homepage Reorganization	Major reorganization of the NCBI homepage with new top-level links to additional databases and services.
02/07	International Database Collaboration	The International Nucleotide Sequence Database Collaboration page describes current projects and provides links to the sites.
01/30	NCBI Structure Group	The NCBI Structure Group (Steve Bryant) has a new page providing access to their structure research, the PKB and MMDB databases, and threading software.

Revised: June 6, 2002.

Volume 12 Number 1 1984 Nucleic Acids Research

A comprehensive set of sequence analysis programs for the VAX

John Devereux, Paul Haeberli* and Oliver Smithies

Laboratory of Genetics, University of Wisconsin, Madison, WI 53706, USA

Received 18 August 1983

ABSTRACT

The University of Wisconsin Genetics Computer Group (UWGCG) has been organized to develop computational tools for the analysis and publication of biological sequence data. A group of programs that will interact with each other has been developed for the Digital Equipment Corporation VAX computer using the VMS operating system. The programs available and the conditions for transfer are described.

INTRODUCTION

The rapid advances in the field of molecular genetics and DNA sequencing have made it imperative for many laboratories to use computers to analyze and manage sequence data. UWGCG was founded when it became clear to several faculty members at the University of Wisconsin that the there was no set of sequence analysis programs that could be used together as a coherent system and be modified easily in response to new ideas.

With intramural support a computer group was organized to build a strong foundation of software upon which future programs in molecular genetics could be based. This initial project has been completed and the resulting programs, written in Fortran 77, are available for VAX computers using the VMS operating system. Most of the programs can be used with only a terminal, although several require a Hewlett Packard plotter.

UWGCG software has been installed for testing at eight different institutions. A simple method has been developed for transferring and maintaining this system on other VAX computers.

DESIGN PRINCIPLES

UWGCG program design is based on the "software tools" approach of Kernighan and Plauger(1). Each program performs a simple function and is easy to use. The programs can be used independently in different combinations so that complex problems are solved by the use of several programs in succession. New programming is simplified since less effort is required to bridge a gap between existing programs.

UWGCG software is designed to be maintained and modified at sites other than the University of Wisconsin. The program manual is extensive and the source codes are organized to make modification convenient. Scientists using UWGCG software are encouraged to use existing programs as a framework for developing new ones. Our copyright can be removed from any program modified by more than 25% of our original effort.

PROGRAMS AVAILABLE FROM UWGCG

The programs described below are named and defined individually in Table 1. Program names in the text are underlined.

Comparisons

Comparisons may be done with "dot plots" using the method of Maizel and Lenk(2). Optimal alignments can be generated by the methods of Needleman and Wunsch(3), of Sellers(4), and the "local homology" method of Smith and Waterman(5). The Smith and Waterman alignment algorithm is also the most sensitive method available for identifying similarities between weakly related sequences.

Mapping and Searching

Mapping is available in several formats. Graphic maps display all of the cuts for each restriction enzyme on parallel lines. This graphic map facilitates selection of enzymes for isolating any region of a sequenced DNA molecule. Sorted maps in tabular format arrange the fragments from any digestion in order of molecular weight to show which fragments are similar in size and thus likely to be confused in gels. Another frequently used mapping format, designed by Frederick Blattner(6), displays the enzyme cuts above the original DNA sequence. Both strands of the DNA and all six frames of translation are shown.

All mapping programs will search for user-specified sequences, allowing features to be marked at the appropriate position on a restriction map. The mapping and searching programs can be used to aid site-specific mutagenesis experiments by showing where mutations could generate new restriction sites. All of the positions in a sequence where a synthetic probe could pair with one or more mismatches can also be located. Sequences related to less precisely defined features such as promoters or intervening sequence splice sites, can be located with a program that uses a consensus sequence as a probe. The

Table 1

Programs Available from UWGCG

Name	Function
DotPlot+	makes a dot plot by method of Maizel and Lenk(2)
Gap	finds optimal alignment by method of Needleman and Wunsch(3)
BestFit	finds optimal alignment by method of Smith and Waterman(5)
MapPlot+	shows restriction map for each enzyme graphically
MapSort	tabulates maps sorted by fragment position and size
Map	displays restriction sites and protein translations above and below the original sequence(Blattner,6)
Consensus	creates a consensus table from pre-aligned sequences
FitConsensus	finds sequences similar to a consensus sequence using a consensus table as a probe
Find	finds sites specified interactively
Stemloop	finds all possible stems (inverted repeats) and loops
Fold*	finds an RNA secondary structure of minimum free energy by the method of Zuker(7)
CodonPreference+	plots the similarity between the codon choices in each reading frame and a codon frequency table(8)
CodonFrequency	tabulates codon frequencies
Correspond	finds similar patterns of codon choice by comparing codon frequency tables (Grantham et al,9)
TestCode+	finds possible coding regions by plotting the "TestCode" statistic of Fickett(10)
Frame+	plots rare codons and open reading frames(8)
PlotStatistics+	plots asymmetries of composition for one strand
Composition	measures composition, di and trinucleotide frequencies
Repeat	finds repeats (direct, not inverted)
Fingerprint	shows the labelled fragments expected for an RNA fingerprint
Seqed	screen oriented sequence editor for entering, editing and checking sequences
Assemble	joins sequences together
Shuffle	randomizes a sequence maintaining composition
Reverse	reverses and/or complements a sequence
Reformat	converts a sequence file from one format to another
Translate	translates a nucleotide into a peptide sequence
BackTranslate	translates a peptide into a nucleotide sequence
Spew	sends a sequence to another computer
GetSeq	accepts a sequence from another computer
Crypt	encrypts a file for access only by password
Simplify	substitutes one of six chemically similar amino acid families for each residue in a peptide sequence
Publish	arranges sequences for publication
Poster+	plots text (for labelling figures and posters)
OverPrint	prints darkened text for figures with a daisy wheel printer

+ requires a Hewlett Packard Series 7221 terminal plotter
* Fold is distributed by Dr. Michael Zuker not UWGCG.

mapping programs can also be used on protein sequences to identify the peptides resulting from proteolytic cleavage.

Secondary Structure

Three programs are available to examine secondary structure in nucleic acids. The program StemLoop identifies all inverted repeats. An implementation of Dr. Michael Zuker's Fold program(7) finds an RNA secondary structure of minimum free energy based on published values of stacking and loop destabilizing energies. The "dot plot" comparison (mentioned above) of a sequence compared to its opposite strand gives a graphic picture of the pattern of inverted repeats in a sequence.

Analysis of Composition and the Location of Genetic Domains

Regions of a sequence with non-random base distribution can be displayed with three graphic tools designed to identify genetic domains. The program CodonPreference(8) identifies potential coding regions by searching through each reading frame for a pattern of preferred codon choices. The CodonPreference plot predicts the level of translational expression of mRNAs and helps identify frame shifts in DNA sequence data. Patterns of codon choice can be compared with the program Correspond(9). When a strong pattern of codon preferences is not expected, the "TestCode" statistic of Fickett(10) can be plotted to show regions of compositional constraint at every third base. Another program plots asymmetries of composition by strand. Strand asymmetries have been associated with genetic domains by several authors(11)(12). A fourth program called Frame marks the positions of rare codons and open reading frames on a graph showing all six reading frames.

Several tools are available to measure content and to count dinucleotide, trinucleotide, neighbor and repeat frequencies. A program that predicts RNA fingerprint patterns and another that tabulates codon frequencies complete the group of programs that analyze composition.

Sequence Manipulation

Sequences may be entered, assembled, edited, reversed, randomized, reformatted, translated, back-translated, documented, transferred, or encrypted rapidly with a large set of sequence manipulation tools.

A screen-oriented editor is available that allows sequences to be entered and checked. After a sequence is entered, it may be reentered for proofreading. Whenever a reentered base is at variance with the original, the terminal bell rings and the position is marked. Existing sequences can be edited quickly by moving directly to a sequence position specified by either a coordinate or a sequence pattern. The program can reassign the terminal's keys to place G, A, T and C conveniently under the fingers of one hand in the same order as the lanes of a sequencing gel.

Programs are available for changing sequence file format. Sequence data from any source can be used in UWGCG programs, and sequence files maintained with UWGCG software can be converted for use in other non-UWGCG programs. For instance, the programs of Roger Staden(13) or Intelligenetics Inc.(14) could be used to assemble a sequence from the sequences of many small sub-fragments generated by DNAase I digestion. The assembled sequence could then be reformatted for use in any UWGCG program. A program is available that transfers sequences to and from other computers.

Sequence Publication

A program, Publish, will format sequences into figures. Publish has alternatives for line size, numbering, scaling, translation and comparison to other sequences. Poster is a program that will plot text on figures.

GENERAL FEATURES OF UWGCG SOFTWARE

Interactive Style

Each program is run by simply typing its name. Every parameter required by the program is obtained interactively. Questions are answered with a file name, a yes, a no, a number, or a letter from a menu. Default answers are displayed. Programs are insensitive to absurd answers and will ask the question again if, for instance, you name a file that does not exist or if you use a nonnumeric character when typing a number. Special features such as plotting features oriented to publication, are obtained by using an extra word next to the program's name when the program is run. Thus parameter queries are kept to a minimum for the normal use of each program.

Data

Both the NIH-GenBank(15) and the EMBL(16) nucleotide sequence data libraries are available "on-line" to any UWGCG program. A Search utility will locate sequences in the libraries by key word. A Find utility will locate library entries containing any specified sequence. A program is available that installs the new data sent periodically from GenBank and EMBL to update their data libraries.

All of the data in the system are stored in text files that can be read and modified easily. Every data file has an English heading describing the contents. The data files may be copied by each user for analysis or modification. Programs recognize and read user-modified input data automatically. Data files can be modified with any text editor.

Sequence File Structure

Sequences are maintained in files that allow documentation and numbering both above and within the sequence. This file format is compatible with both of the nucleic acid sequence libraries and has been adopted as the standard sequence file format by the data base project at the European Molecular Biology Lab. Because genetic manipulations commonly involve linking several molecules of known sequence, UWGCG sequence files are designed to support concatenation by allowing comments to appear within the sequences at any location. Coding sequences or the boundaries between cloning vector and insert, for instance, can be marked within the sequence itself for immediate identification.

Sequence Symbols

All possible nucleotide ambiguities and all standard one-letter amino acid codes are part of the UWGCG symbol set that includes all alphabetic characters plus five additional characters. The proposed IUB-IUPAC standard nucleotide ambiguity symbols(17) are used for the mapping, searching and comparison programs. Lower case characters are used in sequences to indicate uncertainty as distinct from ambiguity. This allows the entire lexicon of symbols to be reused with same meaning, but with the prefix "maybe-." This reuse of the symbol set in lower case makes the uncertainty symbols more complete, understandable and visible.

Symbol Comparison

Sequence analysis programs generally make comparisons between sequence symbols (bases or amino acids) in order to find enzyme sites, create alignments, locate inverted repeats etc. These symbol comparisons are handled in several ways.

Symbol comparisons for alignment, comparison and secondary structure analysis are made by looking up a value in a symbol comparison table for the quality of the match. The table might contain 1's for matches and 0's for mismatches. If amino acids are being compared, however, a real number could be assigned at each position based on some previously assigned chemical similarity of the pair of residues or on the mutational distance between their codons. Standard symbol tables are provided by UWGCG, but the system is designed to allow each user to specify his own values.

Symbols comparisons for mapping and searching operations in nucleic acids are made by converting the IUB-IUPAC symbols into a binary code. The bits of this code represent G, A, T and C with ambiguity symbols causing more than one bit to be set. A group of library functions identify overlap between the bits for each IUB-IUPAC symbol.

Documentation

Documentation is available both in printed form and on the terminal screen. A 350 page manual describes the operation of each program in detail, gives practical considerations and shows what will appear on the screen during a session with the program. Output files and plots are shown for the session. The data for the session shown in the documentation are included with the system so that the each program's operation can be checked. The "on-line" documentation is the same as the manual, but can be changed immediately when a program is modified.

All programs write output to files that are completely documented and sensibly organized for input to other programs. The input data, the program and the parameters used are clearly identified in every output file.

Procedure Library

UWGCG programs are written largely as calls to a library of 250 procedures designed to manipulate biological sequences. These procedures use data and file structures which have been designed to simplify program modification. For instance, standard operations such as reading sequences from files are always handled by a single library procedure. Thus a change in sequence file format requires only one subroutine to be modified for the new format to be acceptable to all of the programs in the system. Command procedures are available to help modify the library. The procedure library can be used by programs written in any language.

DISTRIBUTION OF UWGCG SOFTWARE

Intent

The intent of UWGCG is to make its software available at the lowest possible cost to as many scientists as possible.

Fees

A fee of $2,000 for non-profit institutions or $4,000 for industries is being charged for a tape and documentation for each computer on which UWGCG software is installed. While no continuing fee is required, UWGCG software, like the field it supports, is changing very rapidly. A consortium of industries and academic laboratories is planned to support the project in the future. The consortium will entitle its members to periodic updates and to influence the direction of new programming undertaken by UWGCG in return for a pledge of continuing financial support.

Copyrights

UWGCG retains the copyrights to all of its software and UWGCG must be contacted before all or any part of the its software package is copied or transferred to any machine. UWGCG is, however, mandated to provide research tools to help scientists working in the area of molecular genetics and we are glad to see our source codes become the basis of further programming efforts by other scientists. Copyright can be removed for any program modified by more than 25% of its original effort.

Tape Format

The UWGCG package is usually distributed in VAX/VMS "backup" format on a 9 track magnetic tape recorded at 1600 bits/inch. The system consists of about 1000 files using about 20,000 blocks at 512 bytes/block. The current versions of the GenBank and EMBL nucleotide sequence data bases are normally included which add another 3,000 files and require another 20,000 blocks.

Upon request UWGCG will make a card image tape of all of the Fortran 77 programs and procedures for reading on computers other than the VAX. The card image tape is usually provided at 1600 bits/inch with 80 characters/record and 10 records/block. Adaptation of UWGCG software to systems other than VAX/VMS may take considerable effort.

Equipment Required

UWGCG programs and command procedures will run on a Digital Equipment Corporation (DEC) VAX computer that is using version 3.0 or greater of the DEC VMS operating system. A tape drive is necessary; a floating point accelerator and a DEC Fortran compiler are helpful, but not required. All programs can be run from a DEC VT52 or VT100 terminal. Seven programs, as noted in table 1, require a Hewlett Packard 7221 terminal plotter wired in series with the terminal. Several utilities support a daisy wheel compatible printer attached to the terminal's pass-through port, however, all programs write output files suitable for printing on any standard device.

Inquiries

Inquiries may be sent to John Devereux at the Laboratory of Genetics, University of Wisconsin, Madison, WI, USA 53706, (608) 263-8970. UWGCG is not licensed to distribute Fold(7), but the UWGCG implementation is available from Michael Zuker, Division of Biological Sciences, National Research Council of Canada, 100 Sussex Drive, Ottawa, Canada, K1A OR6 (613) 992-4182.

ACKNOWLEDGEMENTS

UWGCG was started with software written for Oliver Smithies' laboratory with NIH support from grants GM 20069 and AM 20120. UWGCG is directed by John Devereux and is operated as a part of the Laboratory of Genetics with the advice of a steering committee consisting of Richard Burgess, James Dahlberg, Walter Fitch, Oliver Smithies and Millard Susman. UWGCG is currently supported with intramural funds and with fees paid by the faculty and industries using the facility in Madison. This article is paper number 2684 from the Laboratory of Genetics, University of Wisconsin.

*Current address: Silicon Graphics Inc., 630 Clyde Court, Mountain View, CA 94043, USA

REFERENCES

1. Kernighan, B.W. and Plauger, P.J. (1976) Software Tools, Addison-Wesley Publishing Company, Reading, Massachusetts.
2. Maizel, J.V. and Lenk, R.P. (1981) Proceedings of the National Academy of Sciences USA 78, 7665-7669.
3. Needleman, S.B. and Wunsch, C.D. (1970) Journal of Molecular Biology 48, 443-453.
4. Sellers, P.H. (1974) SIAM Journal on Applied Mathematics 26, 787-793.
5. Smith, T.F. and Waterman, M.S. (1981) Advances in Applied Mathematics 2, 482-489.
6. Schroeder, J.L. and Blattner, F.R. (1982) Nucleic Acids Research 10, 69-84, Figure 1.
7. Zuker, M. and Stiegler, P. (1981) Nucleic Acids Research 9, 133-148.
8. Gribskov, M., Devereux, J. and Burgess, R.R. "The Codon Preference Plot: Graphic Analysis of Protein Coding Sequences and Gene Expression," submitted to Nucleic Acids Research.
9. Grantham, R. Gautier, C. Guoy, M. Jacobzone, M. and Mercier R. (1981) Nucleic Acids Research 9(1), r43-r74.
10. Fickett, J.W. (1982) Nucleic Acids Research 10, 5303-5318
11. Smithies, O., Engels, W.R., Devereux, J.R., Slightom, J.L., and S. Shen, (1981) Cell 26, 345-353.
12. Smitb, T.F., Waterman, M.S. and Sadler, J.R. (1983) Nucleic Acids Research 11, 2205-2220.
13. Staden, R. (1980) Nucleic Acids Research 8, 3673-3694.
14. Clayton, J. and Kedes, L. (1982) Nucleic Acids Research 10, 305-321.
15. The GenBank(TM) Genetic Sequence Data Bank is available from Wayne Rindone, Bolt Beranek and Newman Inc., 10 Moulton Street, Cambridge, Massachusetts 02238, USA.
16. The EMBL Nucleotide Sequence Data Library is available from Greg Hamm, European Molecular Biology Laboratory, Postfach 10.2209, Meyerhofstrasse 1, 6900 Heidelberg, West Germany.
17. Personal communication from Dr. Richard Lathe, Transgene SA, 11 Rue Humann, 67000 Strasbourg, France.

Appendix F

DCV-0172

6A

Volume 12 Number 2 1984 . Nucleic Acids Research

Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs

Marilyn Kozak

Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA 15260, USA

Received 15 November 1983; Accepted 21 November 1983

ABSTRACT

5'-Noncoding sequences have been tabulated for 211 messenger RNAs from higher eukaryotic cells. The 5'-proximal AUG triplet serves as the initiator codon in 95% of the mRNAs examined. The most conspicuous conserved feature is the presence of a purine (most often A) three nucleotides upstream from the AUG initiator codon; only 6 of the mRNAs in the survey have a pyrimidine in that position. There is a predominance of C in positions -1, -2, -4 and -5, just upstream from the initiator codon. The sequence CC$\genfrac{}{}{0pt}{}{A}{G}$CCAUG(G) thus emerges as a consensus sequence for eukaryotic initiation sites. The extent to which the ribosome binding site in a given mRNA matches the -1 to -5 consensus sequence varies: more than half of the mRNAs in the tabulation have 3 or 4 nucleotides in common with the CCACC consensus, but only ten mRNAs conform perfectly.

INTRODUCTION

Two years ago I prepared a compilation of the then-available 5'-noncoding sequences of eukaryotic mRNAs (Curr. Topics Microbiol. Immunol. 93, 81-123, 1981). Apart from a bevy of globin and histone mRNAs, only 32 other cellular mRNA sequences were known at that time. In contrast, there are 166 cellular mRNAs in the present compilation, not counting the globins and histones. I have excluded the mRNAs of lower eukaryotes and viruses, only to keep the survey manageable. One of my objectives was to determine whether certain patterns noted in the earlier compilation would be evident with this larger, more diversified set of sequences.

A few points about the selection and presentation of the sequences require explanation. In cases where numerous representatives of a gene family have been sequenced, I have omitted many and chosen those in which the leader sequences show the most divergence. There are exceptions, however. It seemed useful to include certain pairs of mRNAs in which the leader sequences show extensive homology *except near the AUG initiator codon* (e.g. human versus rat preproinsulin). The opposite pattern is also provocative; i.e., sequence conservation *only* near the AUG codon, as in human versus rat immunoglobulin E. Upon inspecting the completed compilation, only two families of mRNAs appeared to be excessively represented: histones and globins. The 5'-noncoding sequences of histone mRNAs are sufficiently varied that they pose little danger of distorting the search for homology among ribosome binding sites. This is less true of the globin sequences, and I have controlled for this as described later in the text.

Nucleotide sequences determined by analyzing genomic DNA have been included only when there is sufficient supplementary data to identify introns that lie upstream from the AUG codon (or to verify their absence) and to estimate the location of the 5'-end of the mRNA. The 5'-end of each mRNA in the table was identified according to one of the following criteria:

(a) Direct sequence analysis of the purified mRNA.
(b) Primer-extension and/or mapping with a single-strand specific nuclease, such as S1. With these techniques there is often a 2- to 4-nucleotide ambiguity in pinpointing the cap site.
(c) Termination of the longest cDNA clone. When the cDNA clone is known to stop significantly short of the 5'-end of the mRNA, the sequence in the table is preceded by an ellipsis (...).
(d) Sequence homology with the corresponding gene from a closely-related species in which the 5'-terminus of the mRNA has been mapped.
(e) Presence in the genomic DNA sequence of an appropriately-positioned TATA box, 25- to 30-nucleotides upstream from the presumptive cap site. In the absence of other supporting data this criterion is rather weak.

The following criteria, identified by code letters in the rightmost column of the table, were used to identify the AUG initiator codon in each message:

(a) The nucleotide sequence corresponds to the known N-terminal amino acid sequence of the *primary* translation product. In some cases amino acid and nucleotide sequence data were derived from two different but related organisms.
(b) The N-terminal amino acid sequence has been determined only for the *mature* protein, which is known (or presumed) to derive from a precursor that carries an N-terminal extension (the "signal peptide") of 15 to 30 amino acids. The indicated AUG triplet is the only candidate initiation site compatible with the synthesis of such a precursor.
(c) The nucleotide sequence has a single open reading frame which either corresponds in size to the known molecular weight of the encoded protein, or includes peptides that are known to be present in the mature protein.
(d) The indicated AUG triplet occurs at the beginning of the longest open reading frame, but the exact size of the primary translation product is not available for comparison. This criterion is rather weak.
(e) The initiation site was deduced from sequence homology with the corresponding gene from a closely-related species in which the start site has been defined.
(f) Under conditions that allow formation of initiation complexes in vitro, the indicated AUG triplet was protected by ribosomes against nuclease digestion.

In 13 of the mRNAs in the table the functional initiator codon has not been definitively identified; the structure of the encoded protein is compatible with initiation at either of two nearby AUG triplets. In such cases I have *predicted* which AUG is most likely to be the (major) initiation site. Those entries are marked with an asterisk in the rightmost column. The AUG initiator codon was predicted based on position (i.e., proximity to the 5'-end of the mRNA) and conformity to

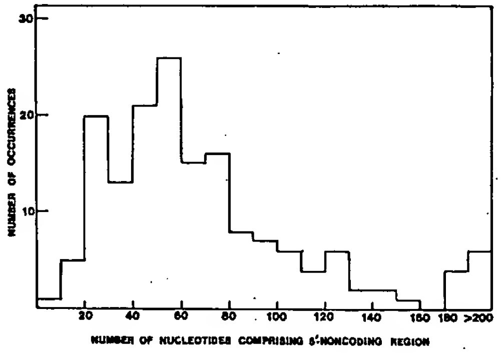


Figure 1. Length distribution of the 5'-noncoding portion of eukaryotic mRNAs. To avoid weighting the distribution by the large number of globin mRNAs in the sequence table, I scored the globin mRNAs only once in this tally.

the consensus sequence CC<u>A</u>CC<u>AUG</u>. In a later section of the text I will explain in greater detail how this was done.

DISCUSSION

A few generalizations emerge from inspection of the sequences tabulated herein.

(i) The length of the 5'-noncoding region varies widely--from 3 to 572 nucleotides. However, 70% of the leader sequences are clustered in the 20- to 80-nucleotide range, as shown in Figure 1. The unusually long leader sequences occur on unusually interesting mRNAs (epidermal growth factor, oncogenes, heat shock proteins), inviting speculation that the structure of the 5'-noncoding region participates in the regulated expression of those genes.

(ii) Translation begins at the 5'-proximal AUG triplet in 95% of the mRNAs tabulated herein. There are only ten mRNAs listed in which one or more AUG triplets occur upstream from the recognized initiation site. The number of "nonfunctional" upstream AUG codons in each of those messages is shown in parentheses at the right edge of the table. [The upstream AUG's are called "nonfunctional" because there is as yet no evidence that ribosomes recognize those sites, but theory predicts that ribosomes should initiate (inefficiently) at the upstream AUG triplets as well as at the AUG codon that heads the long open reading frame.] The number of mRNAs with upstream AUG triplets would increase to 15 if my predictions are correct about which AUG is the major site of initiation in entries 80, 141, 146, 179 and 205. I have dealt elsewhere with the question of how ribosomes get past

ENTRY NO.[1]	MESSENGER RNA	SEQUENCES FROM THE 5'-NONCODING PORTION OF EUKARYOTIC CELLULAR mRNAs[2]	CAP SITE[3]	AUG[4]
001	Acetylcholine receptor pre-α-subunit (Torpedo)	[186]...GTTATTAGAAGTGGCAGATTTGCTTGAAAAGCCAATTATTGAAAGCTGAAGAATGATTCTG	c	b
002	Acetylcholine receptor pre-γ-subunit (Torpedo)	[119]...CCCTCACCCAACGACTCACTCACAAGCTGAGCTAGCACACTCACTGACGACCATGGTACTG	c	b (1)
003	Acetylcholine receptor pre-δ-subunit (Torpedo)	[456]...GACTGCAAATGCTATACTGCAACCAGTAAATTCGGACCTGAGCTCTTTCAAAATGGGGAAC	c	b (8)
004	α_1-Acid preGlycoprotein (rat)	...CTCTTCCTGGGCCGGTGCCTCTGAGTGTCTTCGGCATGGCGCTG	c	*
005	α-Actin, skeletal muscle (chicken)	[73]...GAGGCGTCTCTGCCAGCGGCCCGACGCGCAGTCAGCACAG/ACAGCCAGCAACATGTGTGAC	b	a
006	α-Actin, skeletal muscle (human)	[103]...GGCCCGAGCCGAGAGTAGCAG/TTGTAGCTACCCGCCCAGAAACTAGACACAATGTGCGAC	c	a
007	α-Actin, skeletal muscle (rat)	[71]...CTTCCTACCCTCGGCACCAGGGCCAGAGTCAGAGCAGCAG/AAACTAGACACCATGTGTGAC	b	a
008	β-Actin, cytoplasmic (rat)	[80]...ACAACCTCCTTGCAGCTCCTCCGTCCCGGTCCACACCCGCCACCAG/TTCGCCATGGATGAC	b	a
009	Actin, gene 79B (Drosophila)	[~150]...CTGTTGTACCCTTGTACCCTTGTGTACCGCCCCGCACCAAACTAACCAAACATGTGTGAC	b	a
010	Actin, gene 88F (Drosophila)	[188]...TGGAGCTAACCGTGTGCACTTCCATCTCCCTTCCAGATAAACAACTGCCAAGATGTGTGAC	b	a (3)
011	Alcohol dehydrogenase adult form (Drosophila)	[123]...TACTTAATTGATCAA/ATCGAAAGAGCCTGCTAAAGCAAAAAAGAAGTCACCATGTCGTTT	b	a
012	pre-α-Amylase (barley)	[96]...AAGAAAAGGAGTGTTCTGTACTGTAAGTGAGACAATCGACAGTAGCGCGCGCCATGGGGAAG	c	c
013	pre-α-Amylase, pancreatic (mouse)	[17] GACAACTTCAAAGCAAAATGAAGTTC	a,b	a
014	pre-α-Amylase salivary -- (mouse) liver --	[95]...GCACATG/AAATAAATTAGTTGTTAGAAAGAATACTGCCAACAGCATAGCAAAATGAAATTC [205] CCAAATG/	b	a (1) (1)
015	preAngiotensinogen (rat)	[61]...GCTTGTCTGGGCTGGAGCTAAAGGACACACAGAAGCAAGTCCACAGATCCGTGATGACTCCC	c	b
016	preAntithrombin III (human)	[~67] ...GATCACACTATCTCCACTTGCCCAGCCCTGTGGAAGATTAGCGGCCATGTATTCC	b	b,c
017	preproCalcitonin (rat)	[68]...CATCAGGACCCCGCAGTCTCAGCTCCAAGTCATCGCTCACCAG/GGAGGCATCATGGCTTT	c	c
018	Calmodulin (chicken)	[91]...CGCACCGTGCAGCCGTTAGCTCCCACCGCACCCCGCAGCCGGGCCGAGCCACCATGGCTGAT	c	a
019	pre ε-Casein (mouse)	...TAGGAAGCAAGGACTCATTAATCATGAAGTTC	c	e

020	pre α-Casein (rat)	[61]...GATCATCTCCCAGCTTCTCTCACCCTACTCTTGGGTTCAAGATCTTAGCAACCATGAAACTT	c	a
021	pre β-Casein (rat)	[52] ATCCTCTGAGCTTCATCTTCTCTCTTGTCCTCCGCTAAAGGACTTGACAGCCATGAAGGTC	b	a
022	pre γ-Casein (rat)	[56]ATCATCATCTACCTATTCCTCTCGTCTTCCACTTGGGAAGCAAG/GATCAAGTAACCATGAAGTTC	b,e	a
	preChorion proteins (silkworm)			
023	-m1911	...CTGAATATCCAGCATCATGTTTACC	c	a,c
024	-gene 292a, A-family	[35] ATCATTCTAGATTCAGCAGAACTGTTGTGATCAATATGTCTACT	b	b
025	-gene 18b, A-family	[35] GTCATTCTGAATTTAATATCTCATTCAGTGCAAACATGTCTACC	b	b
026	-gene 401a, B-family	[30] ATCATTCTCAGCTTTGATTTCAAAGTAAACATGAACACT	b	b
027	-gene 10a, B-family	[32] ATCATTATTGAGTTTCCTCTCATACAAACAAAATGGCAGCC	b	b
028	Chorionic gonadotropin pre-β-subunit (human)	...AGACAAGGCAGGGGACGCACCAAGGATGGAGATG	c	a
029	preproChymosin (bovine)	[25] AGCAGCGGCTGGACCCAGATCCAAGATGAGGTGT	c	b
030	preproCollagen-α2, type I (chicken)	[133]...CCACGTCCGGGGGCTCTGCAACACAAGGAGTCTGCATGTCTAGCAAGTAGACATGCTCAGC	b	c (2)
031	preComplement, C3 (mouse)	[56]GCCTCTGCCCACCCCTGCCCCTTACCCCTTCATTCCTTCCACCTTTTTCCTTCACTATGGGACCA	b	b,c
032	preConalbumin (chicken)	[76]...CTGTGACCAACACCGCTGCTCCCCCTCCTCAACACCCAGCTCCCTGCCCCAACATGAAGCTC	a	a
033	preproCorticotropin releasing factor (ovine)	[>127]...AGCGCTGCGCCTCGCTCACCTGCAGAAGCACCTCAGAAG/CGCCCCCTAACATGCGACTG	c	d
034	preCuticle protein I (Drosph)	[42] ATCAGTCAACGTTCGTTCTCGACCAGACAGAAGTCAGCCAATATGTTCAAG	b	b
035	preCuticle protein III "	[45] ATCAGTCTTAGAAGATTTCTAGTCCGACAATCCACCCAAATCAAAATGTTCAAG	b	b
036	Cytochrome P-450 (rat)	[30] ACTGAAGTCTACCGTGGTTACACCAGGACCATGGAGCCC	b	a
037	Dihydrofolate reductase (mouse)	[115]...TTGACGGCAATCCTAGCGTGAAGGCTGGTAGGATTTTATCCCCGCTGCCATCATGGTTCGA (+ longer forms)	b	a
038	preproElastase I (rat)	[21] GTGGTCTACTCTCTCCACAACATGCTGCGC	b,c	b
039	preproElastase II (rat)	[21] ACAGACATCCACGGACACACCATGATCAGG	b,c	c
040	preproEnkephalin (human)	[I29]...CTGAACCCGGCTTTTCCAATTGGCCTGCTCCATCCGAACAGCGTCAAC/TCCATGGCGCGG	b	c
041	preproEpidermal growth factor (mouse)	[353]...TCAGAGGCTCTCGAGAGGTGCAG$^{G}_{A}$AGGACCTGGAAAGGCAGCTAAATAAAAGATGCCCTGG	b,c	d
042	Fatty acid binding protein (rat liver)	[39] CTGTTGGTGGCAGCTGGGAAAGGAAACCTCATTGCCACCATGAACTTC	c	a
043	pre α-Fetoprotein (human)	[42] ATTGTGCTTCCACCACTGCCAATAACAAAATAACTAGCAACCATGAAGTGG	b	c
044	pre α-Fetoprotein (mouse)	[44] ACATCCCACTTCCAGCACTGCCTGCGGTGAAGGAAC$^{C}_{A}$AGCAGCCATGAAGTGG	b,c	a
045	preFibrinogen, Aα (human)	[>60]...TCCTTTCTTTCAGCTGGAGTGTCCTCAGGAGCCAGCCCCACCCTTAGAAAAGATGTTTTCC	c	*

046	preFibrinogen, γ (rat)	[53]	AGAGGTCACAGTGCTGGCTGTAAGGGGCTGGGAGCACCGGTCACCCAGACACTATGAATTGG	b,e	b
047	preFibroin (silkmoth)	[24]	ATCAGTTCGGTTCCAACTCTCAAGATGAGAGTC	a,b	c
048	preproGastrin (porcine)	[61]	...ACTGAGGCACCAGGCCAACAGCAGCACACCTGCCTCCCAGCTCTGCAGTCAAGATGCAGCGA	c	c (1)
	α-Globin family:				
049	-chicken, embryonic (π')	[55]	ACAACCTGCTCTGGGTGTTCACTGAAGGGAGCCTGAGCCAGCACTCTCCTGCACAATGGCACTG	d,e	a
050	-duck, major adult (α^{A})	[36]	ACCCGTGCTGGGGGCTGCCAACGCGGAGCTGCAACCATGGTGCTG	b	a
051	-duck, minor adult (α^{D})	[42]	ACAGAAACCCGTCAGTTGCCAGCCTGCCACGCCGCTGCCGCCATGCTGACC	b	a
052	-human, adult	[37]	ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTG	a	a
053	-human, embryonic (ζ)	[55]	ACCAAGGCCAGTCCTGAGCAGGCCCAACTCCAGTGCAGCTGCCCACCCTGCCGCCATGTCTCTG	b	a
054	-mouse, adult	[32]	ACTTCTGATTCTGACAGACTCAGGAAGAAACCATGGTGCTC	a	a
055	-rabbit, adult	[36]	ACACTTCTGGTCCAGTCCGACTGAGAAGGAACCACCATGGTGCTG	a	a
056	-Xenopus, major adult		...TGCACAACACAAACAGCAACCATGCTTCTT	c	c
	β-Globin family:				
057	-chicken, adult	[77]	...GAGCCCAGACCTCCTCCGTACCGACAACCACACGCTACCCTCCAACCGCCGCCATGGTGCAC	b	a
058	-duck, adult	[∿60]	...AGCCGAGACCTCCTCCGTACCTGCAGCCACACGCTACCCTCCGCCCGACACCATGGTGCAC	c	a
059	-chicken, embryonic (ρ)	[45]	AAGCTCTGAGTGCTCCCACAGCCGCACGCCAACCCCGCTGCCACCATGGTGCAC	c	a
060	-goat, adult (β^{A})	[52]	ACACTTGCTTCTGACACAACCGTGTTCACTAGCAGCTACACAAACAGACACCATGCTGACT	d,e	a
061	-human, adult	[50]	ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCAC	a	a
062	-human, fetal (γ)	[53]	ACACTCGCTTCTGGAACGTCTGAG$\substack{A\\G}$TTATCAATAAGCTCCTAGTCCAGACGCCATGGGTCAT	c,d	a
063	-human, embryonic (ε)	[53]	ATATCTGCTTCCGACACAGCTGCAATCACTAGCAAGCTCTCAGGCCTGGCATCATGGTGCAT	d	a
064	-mouse, major adult	[52]	ACATTTGCTTCTGACATAGTTGTGTTGACTCACAACCCCAGAAACAGACATCATGGTGCAC	a	a
065	-mouse, embryonic	[52]	ACTTCGCTTCTGACACTCCTGTGATCACCAGCAACCTCCCAGACTTGCCATCATGGTGAAC	d	a
066	-rabbit, adult	[53]	ACACTTGCTTTTGACACAACTGTGTTTACTTGCAATCCCCCAAAACAGACAGAATGGTGCAT	a	a
067	-rabbit, embryonic (β3)	[62]	...TCTGAGACATCTGAGACTATCAGCAAGCTCAGCGAGCTCCTAGACCAGACATCATGGTTCAT	b	a
068	-Xenopus, major adult	[46]	ACTTGTTCTTTTTGCAGAAGCTCAGAATAAACGCTCAACTTTGGCCATGGGTTTG	b	c
069	-Xenopus, larval		...TAGCAGCTACTCCCATCTACAGCCACCATGGTGCAC	c	c
070	pre α_{2u}Globulin (rat)	[∿68]	...CCATCAGCAGAGAGATTGTCCCGACAGAGAGGCAATTCTATTCCCTACCAACATGAAGCTG	c	a
071	preproGlucagon I (anglerfish)	[∿58]	...AGGAACTAACAGCACTATTTGAGGGAGAAAAAGAATAAATACGGTTGTAAACATGAAACGC	c	c
072	preproGlucagon II (")		...GAAGCTCAAACAATGACAAGT	c	c
073	preproGlucagon (hamster)	[103]	...TGCACCTGCTCACCTGCTCTCCGCTCAGTCACAGCAGCAGGCAGAAAAAAAAATGAAGAAC	c	c
	preGlue proteins (Drosophila)				
074	-Sgs-4	[13]	TTCCAAAGTCAAGATGCGCTTG	b	d
075	-Sgs-3	[29]	ATCAGTTTGTGGAGAATTAAGTAAAAAACATGAAGCTG	c,d	d
076	-Sgs-7	[33]	ATCTGGTAAAGTAGTCTCAATCTAGATAGAACCATGAAACTG	b,d	d
077	-Sgs-8	[33]	ATCTGGTAAAGTTAATTTGTTAAAGCAACAACCATGAAGCTG	b,d	d
078	Glycoprotein hormones (human) pre-α-subunit	[100]	...CAGTCAACCGCCCTGAACACATCCTGCAAAAAGCCCAGAGAAAG/GAGCGCCATGGATTAC	d,e	a

079	Glycoprotein hormones (mouse) pre-α-subunit	[∿100]	...AATCACCTGCCCAGACACATCCCTCAAAAGTCCAGAGCTTGCAGAAGAGCTATGGATTAC	c	b
080	preGrowth hormone (bovine)	[61]	...AGGACCCAGTTCACCAGACGACTCAGGGTCCTGTGGACAGCTCACCAGCTATGATGGCTGCA/	d,e	*
081	preGrowth hormone (human)	[60]	...AAGGCCCAACTCCCCGAACCACTCAGGGTCCTGTGGACGCTCACCTAGCTGCAATGGCTACA	b,e	b
082	preGrowth hormone (rat)	[60]	...TCGAGCCCAGATTCCAAACTACTCAGGTCCTGTGGACAGATCACTGAGTGGCGATGGCTGCA	b,e	a,b
	Heat shock (Drosophila):				
083	-70K protein	[247]	...GAGAACTCTGAATACTTTCAACAAGTCGTTACCGAGGAAGAACTCACACACAATGCCTGCT	b,e	c
084	-22K protein	[252]	...ATTTGCATAAGAAGCTTTATTTGGAAAAACCCAAGTTACCTTATCAACTACAATGCGTTCC	b,e	c
085	-23K protein	[112]	...AAGAAGTTTATTCTTTGAAGGAGAATCATCTTGAAGCAATTAAAAAACAAAAATGGCAAAT	b,e	c
086	-26K protein	[181]	...AGAAAAAATATTTCAATCTCGCAAAGGAACATAACCTAAGGAAAAAGTAAAAATGTCGCTA	b,e	c
087	-27K protein	[121]	...AAAAATTCCTTTGTCTAGACAGGGTTGTGAATAAAGAGAAAAAAAATCAAAAATGTCAATT	b,e	c
	Histocompatibility (MHC) antigens:				
088	-Class I: mouse (pre)H-2K^d	[25]	AAGTCGCTAATCGCCGACCAGTGCGATGGCACCC	c,e	b
089	-Class II: mouse(pre)E_α 34K	[48]	GCCTCAGTCTGCGATCGCTTCTGAACCCACCAAACACCCAAGAAGAAAATGGCCACA	b	b
090	- " mouse(pre)I-A_β 29K		...GGCATTACCTGTGCCTTAGAGATGGCTCTG	-	b
091	- " human(pre)DR_β 29K		...CTCCTCTGGCCCCTGGTCCTGTCCTCTTCTCCAGCATGGTGTGT	c	b
	Histones:				
092	-chicken H1, embryonic	[38]	AGTGGTTCCCCCCGATCTGTGGAACGACGTCCGTCACCATGTCGGAG	b	c
093	- " H2A.1, adult	[146]	...GTTGCGCGCTGCGTTCTTGGCTTGTTTGCTCTCTGAGTGTTCAGTCGCTGCGATGTCGGGG	e	a
094	- " H2A.F, embryonic	[70]	...GAGGCGGTTCGAGCGGCGGATTCGGGGACCGGGGCTCGGGCGGCGGCGGCACCATGGCAGGT	c	c
095	- " H2B	[36]	ACTCCGCTGCGCCGAAGGGATCGTGGAGAGTTCGACATGCCTGAG	e	a
096	- " H4, embryonic	[26]	GGTGTCGGACGACTCAGGCTCTCGGCATGTCTGGC	b	a
097	- " H5, erythrocyte	[109]	...CTTTTTTAAGCTCCCTAACCCCAGTGCCCTGCCGTGGGGTGAAGCGGCGGCCATGACGGAG	c,e	a
098	-human H3	[27]	ACTTTCGGTTGCAGAGCAGTTCTGCGAATGGCGCGT	e	a
099	- " H4	[36]	ACAAAGCGTTGGGTGAGACGCCTCTTGCTCGTCGTCATGTCTGGC	e	a
100	-sea urchin, early, H1	[38]	ATTCGTTTTGTTAACCTCCCGACGCACCGTATATCAAGATGGCTGAG	b	
101	(S. purpuratus) H2A	[70]	...TTGTTACGTTCTGAACTTCGCTTCCGATTTATTCTAAACTCATCAACAACATCATGTCTGGC	b	a
102	H2B	[78]	...CTTGACATACTCGTTTCGCTGCATCTTTACAGACCAGAAAACCTCAATTCATCATGGCTCCA	b	a
103	H3	[55]	ATTCATCCCGTCACTCGTATTTGAAGTACTGAATTACCTACTGTCCCAAGCAACTATGGCACGC	b	a
104	H4	[67]	...ATATCCAGTCTACAGGATCACACAGAACTCGCTCTCAACTATCAATCATCATCATGTCAGGT	a,b	a
105	-sea urchin, early H2A	[60]	...CGCTTCGTTCACAACCTCGCTTCGCTCTCAGCTCGTTAACCAACCAACCATCATGTCTGGA	b	a
106	(P. miliaris, h22) H2B	[76]	...TCGTGACACTCGCATCGTTCTGCTCCTAAGACATCAGAAAACTTCATCTCACCATG	b	a
107	H3	[54]	ATTCATCTCGTCACCCTCGTTTGAAACACTGATCCAACTGTCTCCCAATCAACCATGGCACGA	b	a
108	H4	[60]	...AGCGAAACGTCCAGTCGTCAGCATCGCACTAAGACTCTCTCTCAATCTCCATAATGTCAGGC	d	a
109	-sea urchin, late H3	[∿27]	ATCAGTTGACTATCTGAAAAATCAACAATGGCCCGT	b,e	a
110	(L. pictus) H4	[∿23]	ATCTCAAACAGGAAACTATCAACATGTCTGGA	b,e	a
111	-Xenopus H1	[28]	AGTTGGCGTGAGTGAATTTACTTCAAAGATGACAGAA	e	a
112	" H2A	[47]	AGTCTACAACATCTTCTTGATTGTGGTTGATTTGTAGCACAGTAATCATGTCTGGA	e	a
113	" H2B	[35]	ACAGTTTTGTAGGCTGAGAGAGAAGCAGCACAATTATGCCTGAA	e	a
114	" H3	[80?]	...CGGGTTACCCGCTGCCGCAGTTACACGCTTAGGTTTCTTTAACTGATACACTATGGCCCGT	e	a
115	" H4	[28]	ATATTGTGTTTCAAGAGCTCAAGAAAGAATGTCTGGA	e	a

116	Hypoxanthine phosphoribosyl-transferase (HPRT), human	[85]	...GCCACCGGCTTCCTCCTCCTGAGCAGTCAGCCCGCGCGCCGGCCGGCTCCGTTATGGCGACC	c	c
117	HPRT, mouse	[87]	...TCCGCCGGCTTCCTCCTCAGACCGCTTTTTGCCGCGAGCCGACCGGTCCCGTCATGCCGACC	c	c
	Immunoglobulins:				
118	-(pre)κ_{167} light chain, mouse	[29]	ATCAGAACTCTTCACACCAAGTTCTCAGAATGAGGTGC	b	b
119	-(pre)κ light chain, mouse	[3; 18]	GAAATGCATCACACCAGCATGGCATC	a,b	a^+
120	-(pre)λ_I light chain, mouse	[40]	GCTGACCAATATTGAAAAGAATAGACCTGGTTTGTGAATTATGGCCTGG	b	a
121	-(pre)heavy chain, mouse	[45]	TGTCCCAATCTTCACATTCAGAAATCAGCACTCAGTCCTGTCACTATGAAGTTG	c	b
122	-(pre)heavy chain, mouse	[30]	ATCCTCTTCTCATAGAGCCTCCGTCAGAGCATGGCTGTC	b	c
123	-(pre)heavy chain, human IgE	[56]	CACCCAGCAACCACATCTGTCCTCTAGAGAATCCCCTGAGAGCTCCGTTCCTCACCATGGACTGG	c	b
124	-(pre)heavy chain, rat IgE	[71]	...TGAACGCACAAATCTTTTCCTTCAGTGACTAGCACTGATACAGAATATTCACCATGGACTTG	c	b
125	preproInsulin (chicken)	[54]	AGCCTGATGAATAAAATATTCCTTTCCTCTTCAGAAG/GCCTCCCCCAGCTCATCATGGCTCTC	d	b (1)
126	preproInsulin (human)	[59]	...CAGGACAGGCTGCATCAGAAGAGGCCATCAAGCAG/ATCACTGTCCTTCTGCCATGGCCCTG	d	b
127	preproInsulin (rat-I)	[57]	...TAAGTGACCAGCTACAATCATAGACCATCAGCAAGCAG/GTCATTGTTCCAACATGGCCCTG	b,e	a
128	preproInsulin (anglerfish)	[85]	...CTCTTCATCAACTCATTCATCACCTCCTCCTACTCTACAG/TTCTACTGCAGCATGGCGGCT	c	a
129	preproInsulin (hagfish)	[49]	CCAACTTGTCGCACTAGAGGGACGGAGACCATTATCCAAAAGCAACCAAATGGCGCTC	c	b
130	preproInsulin (salmon)	[>71]	...ATTCTTCTACAGAGCTTCTTCACCACCACCAACTCCATAGCCTACCATCACCATGGCCCTC	c	b
	Interferons:				
131	-pre-α (LeIF-A), human	[67]	...AAGGTTTAGGCTCACCCATTTCAACCAGTCTAGCAGCATCTGCAACATCTACAATGGCCTTG	c,e	b
132	-pre-α (LeIF-D), human	[67]	...AAGGTTCAGAGTCACCCATCTCAGCAAGCCCAGAAGTATCTGCAATATCTACGATGGCCTCG	c,e	b
133	-pre-α (LeIF-C), human	[67]	...CAAAGTTCAAGGTTACCCATCTCAAGTAGCCTAGCAACATTTGCAACATCCCAATGGCCCTG	b	b
134	-pre-α (LeIF-H), human	[69]	...AGGTTCAGTGTTACCCCTCATCAACCAGCCCAGCAGCATCTTCAGGGTTCCCAATGGCATTG	b	b
135	-pre-α (leukocyte), mouse	[67]	...GAAGGATCAGGACCAAACAGTCCAGAAGACCAGAAGCTTTGGCAACACTCACCATGGCTAGG	d,e	b
136	-pre-β (fibroblast), human	[75]	...AAGCCTTTGCTCTGGCACAACAGGTAGTAGGCGACACTGTTCGTGTTGTCAACATGACCAAC	b,c	b
137	-pre-β (fibroblast), mouse		...ACACCAGCCTGGCTTCCATCATGAACAAC	c	b
138	-pre-γ (immune), human	[128]	...TGATACAAGAACTACTGATTTCAACTTCTTTGGCTTAATTCTCTCGGAAACGATGAAATAT	b,e	d
139	preInterleukin-2, human	[47]	ATCACTCTCTTTAATCACTACTCACAGTAACCTCAACTCCTGCCACAATGTACAGG	b,c	*
140	preproKallikrein, mouse	[42]	AGCTCCAAGCTCACTGCCTGCTGCTCCTGAACACCTGTTACCATGTGGTTC	b	b
141	preproKallikrein, rat		...AAGCTCAGCACCTGCTGCTCCTGCATGCCTGTTACCATGTGGTTC	c	*
142	Keratin, epidermal (mouse)		...CCGACTCCTGGGCCACACCACACCACCATGTCTGTC	c	b
143	Keratin, B2A (sheep)	[50]	ACTCAAAAACTTCTCTTAACAACCCTCCTCTCAACCCAACTCCTGACACCATGGCCTGC	e	a
144	preKininogen, low MW (bovine)	[133]	...CCCAACTGGCTCCCAATTCCTGGTGAAACTCCTAAAGGGAGACTGTTAAATCATGAAATTA	c	c
145	pre-α-Lactalbumin, guinea pig	[84]	...TTCTGGGTGTATGGGAAGAGCAGGGGCATTTCAGGTTCACAGCAGCAGCCAAAATGATGTCC	c	b (2)
146	preLectin (French bean)	[16]	NNATGAATGCATGATCATGGCTTCC	c	*
147	Leghemoglobin (soybean)	[49]	AACTTGCATTGAACAATTAATAGAAATAACAGAAAAGTAGAAAAGAAATATGGGTGCT	b,e	a

148	apoLipoprotein II, VLD, chick	[77]	...GAAAGCAGGACAG/GTCTCTTGGTGTAAAGGGCTGAACTGGTACCAACAAACCATGGTGCAA	b	a
149	apoLipoprotein A-I (human)		...TCCCCCACGGCCCTTCAGGATGAAAGCT	c	a
150	apoLipoprotein E (rat)		...ACTGGCCAATCACAACTGGGAAGATGAAGGCT	c	a
151	preLysozyme (chicken)	[29]	AGTCCCGCTGTGTGTACGACACTGGCAACATGAGGTCT	b	a
152	preproMelittin (honeybee)		...AGCGAATTAACAGAATTAACAGGAAGGAAGGAAGGAAGCGATCGGAGAAATCATGAAATTC	c	a
153	Metallothionein-II (human)	[69]	...CCAGCGAACCCGCGTGCAACCTGTCCCGACTCTAGCCGCCTCTTCAGCTCGCCATGGATCCC	b	a
154	Metallothionein-I (mouse)	[73]	...CTGAGTACCTTCTCCTCACTTACTCCGTAGCTCCAGCTTCACCAGATCTCGGAATGGACCCC	b	a
155	pre-β2-Microglobulin (mouse)	[>52]	...ATTTCAGTGGCTGCTACTCGGCGCTTCAGTCGCGGTCGCTTCAGTCGTCAGCATGGCTCGC	c	a
156	Myoglobin (seal)	[70]	...CAGGACACCGAGTCAGCCGGGACTTGCTCTTCTTGTCTTCCTCAGACTGCACCATGGGGCTC	b	a
157	Myosin, skel. L chain (chick)		...CCTCTCAGCTAATCCCTCCGGCCCGTCCGCTCTACTTTTCCAACTCTCAATCATGTCCTTC	c	c
158	prepro-β-Nerve growth factor (mouse)	[74]	...GCTGGCCTTATATTTGGATCTCCCGGGCAGCTTTTTGGAAACTCCTAGTGAACATGCTGTGC	c	d
	protoOncogenes:				
159	-c-*fos* (human)	[154]	...CCCACCTGTCTCCGCCCCTCGGCCCCTCGCCCGGCTTTGCCTAACCGCCACGATGATGTTC	e	*
160	-c-*myc* (human)	[572]	...AGACGCTGGATTTTTTTCGGGTAGTGGAAAACCAG/CAGCCTCCCGCGACGATGCCCCTC	c	d
161	-c-Ki-*ras*2 (human)	[>181]	...GGGGCCAGAGGCTCAGCGGCTCCCAGGTGCGGGAGAGAG/GCCTGCTGAAAATGACTGAA	c	c
162	-c-*src* (chicken)		...TCGCGTACCACTGTGGCCAGGCGGTAGCTGGGACGTGCAG/CCCACCACCATGGGGAGC	-	c
163	preproOpiomelanocortin (bovine)	[128]	...GGAGCCGCCCGAGGCAGCTTCCCCGTGACAG/AGCCTCAGCCTGCCTGGAAGATGCCGAGA	b	a
164	preproOpiomelanocortin (human)	[107]	...CCCGCCCTCAGAGAGCAGCCTCCCGAGACAG/AGCCTCAGCCTGCCTGGAAGATGCCGAGA	d,e	a
165	Ovalbumin (chicken)	[64]	...AAAGCTGTATTGCCTTTAGCAGTCAAGCTCGAAAG/ACAACTCAGAGTTCACCATGGGCTCC	a,b	a,f
166	preOvomucoid (chicken)	[53]	ATCTCAGGAGCAGAGCACCGGCAGCCGCCTGCAGAGCCGGGCAGTACCTCACCATGGCCATG	b,c	a
167	preproParathyroid hormone (bovine)	[100]	...TCAGCTGCTAATACATTTGAAAGAAGATTGTATCCTAAGACGTGTG/TTAATATGATGTCT	b	a
168	preproParathyroid hormone (human)	[90]	...AGCTACTAACATACCTGAACGAAGATCTTGTTCTAAGACATTGTATG/TGAAGATGATACCT	d,e	a (1)
169	prePepsinogen (human)	[~54]	CTGCACCTTCCTCCCGTCTTGCCTTCTCCCTCGAGTTGGGACCCGGGAAGAACCATGAAGTGG	b,e	b
170	pro(?)Phaseolin	[77]	...CTACTCTACTACTATAATACCCCAACCCAACTCATATTCAATACTACTCTACTATGATGAGA	c	*
171	Phosphoglycerate kinase, human	[80]	...GGCTCCCTCGTTGACCGAATCACCGACCTCTCTCCCCAGCTGTATTTCCAAAATGTCGCTT	c	a
172	prePlacental lactogen, human		...CTGTGGACAGCTCACCTAGCGGCAATGGCTGCA	c	a

173	prePlasminogen activator (human)	[84]...	GCGAGGAAAGGGAAGGAGCAAGCCGTGAATTTAAGGGACGCTGTGAAGCAATCATGGATGCA	c	*
174	preProlactin (rat)	[51]	AGTGGTTCTCTTAGGACTTCTTGGGGAAGTGTGGTCCCAGTGGTCATCACCATGAACAGC	c,e	a
175	preProlactin (bovine)	[67]...	GCCATAGGACGAGAGCTTCCTGGTGAAGTGTGTTTCTTGAAATCATCACCACCATGGACAGC	c	b
176	Protamine (trout)	[14]	ATCCATCAATCACAATGCCCAGA	b	a
177	Pyruvate kinase (chicken)	[80]...	GCTTTGGGCACGGCGGCGGAAGCAGCAGCAGGAGACACCGAACTCCAGTAACCATGTCGAAG	c	c
178	preproRelaxin (rat)	[∿60]	...CTGAACCGCCCAGGAGCACCGCCCAGGAGCAGCCCAGACCGGAATGTCCAGC	c	b
179	preproRenin (mouse)	[55]	TCTGGGCTACACAGCTCTTAGAAAGCCTTGGCTGAACCAGATGGACAGGAGGAGGATGCCTCTC	c	*
180	preRibonuclease, panc. (rat)	[∿75]...	CAATTTGCTCTGGAATTCAAACGTTTAGGCTCCTCAGAGCAGCAAAGCCACTATGGGTCTG	c	b
181	Ribosomal protein S19, Xenopus	[>46]	...AATTCTTTAATCCTTTCTTTTGTCACCGTGAGAGATAGCCGGCAAGATGAATGAC	c	c
182	preRibulose bisphosphate carboxylase (soybean)	[45]	ATCTGGCAGCAGAAAAACAAGTAGTTGAGAACTAAGAAGAAGAAAATGGCTTCC	b	a
183	Seminal vesicle pre-secretory protein IV (rat)	[22]	AGTCAAGAGCTTTTCTGGCAAGATGAAGTCT	b	b
184	pre(?)Sericin (silkmoth)	[54]	ATAGTCGTCTTATCATCGGGTCTCTAAGGATCAAGCGATCCAAAGACCGCCAACATGCGTTTC	b	c
185	Serum preproalbumin, chick	[41]	AGCATTTTTGAATAATTTAGCCCACATCATAATCTGCAGCCATGAAGTGG	b	a
186	Serum preproalbumin, human		...GCTTTTCTCTTCTGTCAACCCCACACGCCTTTGGCACAATGAAGTGG	c	a
187	preproSomatocrinin (human)	[112]...	TGCCTCGGAGCAGAGGGATATCTGCCGCATCAGGTGCCACCCCGGGTGAAGGATGCCACTC	b	c
188	preproSomatostatin-II (anglerfish)	[59]...	CAAACCCAGCAGAACCAGTAGAACCAGCAGAAGACACCAGACCAGCAGACAGTATGCAGTGT	c	c
189	preproSomatostatin-22 (catfish)		...CCGCCCAGCTCGAAATCTTCACCAGCTACCAAGAAGATGTCGTCT	c	c
190	preproSomatostatin-14 "	[114]...	GCCCTCCCTCCACCAATTTTTCCACCAAATCCAGCTTTATTTCTTTTTGAAGATGCCCTCC	c	c
191	preproSomatostatin-I (human)	[105]...	CGCGCCTAGAGTTTGACCAGCCACTCTCCAGCTCGGCTTTCGCGGCGCCGAGATGCTGTCC	c	c
192	preproSomatostatin (rat)	[81]...	CTGCGTCTAGACTGACCCACCGCGCTCAAGCTCGGCTGTCTGAGGCAGGGGAGATGCTGTCC	c	c
193	pre-Steroid binding protein C1, rat prostate	[44]	TGGAAGAGTTCATTTGCTCACCATTTGCTAAGTAGAAAACTGAAATGAGCACC	b	b
194	pre-Steroid binding protein C2, rat prostate	[42]	TGGGAGAGTCCATTTGCTCAGTCTAAAAGCAAACTGAGCACCATGAGGCTG	b	b
195	pre-Steroid binding protein C3, rat prostate	[55]	GAGTTTCCTGATTTCTGTCTTGGACAACAGAACAACCCACAGGGACTGCCTCAACATGAAGCTG	b	b
196	preproThaumatin	[31]	AAAGCGCAGCCTCAATTGGCATCATACATCAATGGCCGCC	b	c
197	Thyrotropin, pre-β-subunit (rat)	[89]...	CGAGTGGAGAGAAAAATATTCTGCTTCAGTGAAGAGCTGGGGTTGTTCAAAGCATGAGTGCT	c	b
198	preproTrypsinogen (rat)	[12]	CCTTCTGCCACCATGAGTGCA	b	b

199	α-Tubulin (rat)	[100]	...AACACCTCCTCCTCGCCTCCGCCATCCACCCGGCAGCCGCGAAGCAGCAACC*ATG*/CGTGA	b	a
200	β-Tubulin (chicken)	[87]	...AGAGCGGGAGGTGACGGAGCGGGAGCAGCGCGGCACCGGCAGACACCGGCATC*ATG*CGTGAG	c	a
201	β-Tubulin (human)	[159]	...TTTTCTTGCCCCATACATACCTTGAGGCGAGCAAAAAAATTAAATTTTAACC*ATG*AGGGAA	b	a
202	preUteroglobin (rabbit)	[47]	AGATCACCGGATCCAGAGCCAGCCCAGAGCCTTCCCATTCTGCCACC*ATG*AAGCTC	b,c	a
203	preproVasoactive intestinal polypeptide (human)		GGGGAGCACGACTGGGCGAGAGGCACAGAA*ATG*GACACC	b	c
204	preproVasopressin neurophysin II (bovine)	[49]	GCACAGTCTACAGAGCAGCACTGCGCACGTGTGCCCACGCGTGCCAGG*ATG*CCCGAC	b	b
205	preproVasopressin neurophysin (rat)		AGCAGAGCAGAGCTGCACGCAGTGCCCACCT*ATG*CTCGCC*ATGATG*CTCAAC	b	*
206	Vitellogenin II (chick)	[13]	ATTCACCTTCGCT*ATG*AGGGGG	b	d
207	Vitellogenin I (Drosophila)	[≥58]	...ACTCACTCAGTGTTGAAGTCGCATCCGCAGGACCAAATCCCAAATCCGAACC*ATG*AACCCC	b	*
208	Vitellogenin II (Drosophila)	[≥51]	*ATG*CAGTACAATTTGGTACGGTGTCTGAAAAAGTCGAACTTGGAAGCCACA*ATG*AATCCT	b	*
209	preWhey acidic protein (mouse)	[26]	ATCAGTCACTTGCCTGACCCGGTACC*ATG*CGTTGC		c
210	preZein, 19K (maize)	[57]	...CACATATTATTGAGACCAACTAGCAATATAGAAAGCACAATATTGTACCAATA*ATG*GCAGCC	c	b
211	preZein, 22K (maize)	[>67]	...TCAGCATTCAAAAACACACCAAGCGAAGCGCACTAGCAAC$^{G}_{A}$ACCTAACA$^{A}_{C}$CA*ATG*GCTACC	c	c

Footnotes:

[1] The number assigned here to each mRNA is used again to identify the corresponding references in the bibliography.

[2] The table shows the sequence of the plus strand of DNA, from which the sequence of mRNA can be derived by substituting U for T. All ATG triplets are shown in italics. The sequences are aligned by using the ATG triplet that is known (see footnote 4) or predicted (see text) to be the functional initiator codon. The positions of introns are indicated by a diagonal line. For mRNAs in which the 5'-noncoding sequence exceeds 56 nucleotides, I have shown only the portion nearest the ATG triplet. *The number in brackets preceding the sequence indicates the (approximate) full length of the 5'-noncoding sequence*, not counting the m^7G cap or the ATG triplet. There is likely to be a 2- to 4-nucleotide uncertainty when S1 nuclease was used to map the cap site. In cases where cDNA clones were analyzed, the 5'-noncoding sequence may be a little longer than indicated. If the missing portion of the leader is suspected to be more than a few nucleotides, no figure is given for the overall length of the 5'-noncoding region.

[3] The criteria (a-e) used to identify the 5'-terminus of each mRNA are summarized in the text.

[4] The criteria (a-f) used to identify the ATG initiator codon are summarized in the text. An *asterisk* in this column means that the ATG triplet used to align the mRNA is *predicted*, but is not known to be the functional initiator codon. I have also marked in this column those mRNAs that have ATG triplets upstream from the functional initiation site. *The number of upstream ATG's is indicated in parentheses*. Entry 119 (marked +) is, to my knowledge, the only cellular mRNA in which two functional initiator codons have been identified: the ATGs in positions 4-6, and 19-21.

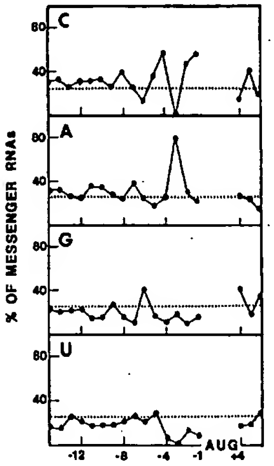


Figure 2. Frequency distribution of each nucleotide around the functional initiator codon in 198 mRNAs listed in the table. The calculations presented here do not include the 13 mRNAs in which the translational start site was predicted but has not been verified. The nucleotide immediately preceding the AUG codon is numbered -1; nucleotides +4 to +6 represent the start of the protein coding sequence. The dotted line across each panel indicates the 25% value that would be expected on a random basis. To ensure that the results were not distorted by the inclusion of too many closely-related globin mRNA sequences, I recalculated the frequency of occurrence of each nucleotide in positions -1 through -8, omitting all of the globin sequences. Although the absolute values changed somewhat (e.g. G in position -6 dropped from 40% to 36%), the rank order of nucleotides in each position remained unchanged.

the upstream AUG triplet(s) in such messages (Kozak, Microbiol. Rev., 47, 1-45, 1983; Kozak, manuscript submitted). The main point to note here is that such mRNAs are rare. The "first-AUG-rule" holds for 93% to 95% of the entries in the table.

(iii) The sequences in the table have been searched manually for signs of a conserved motif that might uniquely identify AUG initiator codons. The most conspicuous conserved feature is presence of a purine (most often A) in position -3; i.e., three nucleotides upstream from the initiator codon. As illustrated in Figure 2, 79% of the mRNAs that were counted have A in that position, 18% have G, and only 3% (a total of 6 messages) have a pyrimidine in position -3. The strong preference for a purine in position -3 is peculiar to AUG triplets that serve as initiator codons. *Pyrimidines* are favored in the -3 position preceding AUG triplets that lie *upstream* from the initiation site, in those rare mRNAs that have upstream AUGs (Kozak, Nuc. Acids Res. 9, 5233-5252, 1981); and the nucleotide frequency in position -3 is almost perfectly random around AUG triplets that code for methionine at *internal* positions in polypeptide chains (Kozak, 1983, op. cit.). Although no other position is as highly conserved as the purine in position -3, the distribution of nucleotides is decidedly nonrandom in every position from -1 through -6, and perhaps beyond. The predominance of C in positions -1, -2, -4 and -5 was evident in an earlier survey (Kozak, 1981, op cit.) and is confirmed here. The preference for G in position +4, noted in the previous survey, is less evident here. From the data in Figure 2, the sequence CC$^{A}_{G}$CCAUG(G) emerges as a consensus sequence for eukaryotic initiation sites. The extent to which a given message matches the -1 to -5 consensus sequence varies considerably: only 10 mRNAs in the table conform perfectly to the CCACC sequence; in more than half of the mRNAs, 3 or 4 of the nucleotides directly preceding the AUG codon match the consensus sequence; about 10% of the mRNAs have a purine in position -3 but otherwise differ entirely from the -1 to -5 consensus. The 6 mRNAs in the tabulation that lack a purine three nucleotides upstream from the initiator codon do not seem to compensate by conforming closely to the other four consensus positions. Recent site-directed mutagenesis experiments (Kozak, manuscript submitted) have confirmed the importance of the purine in position -3, but there is as yet no evidence that cytosine in positions -1, -2, -4 and -5 contributes to recognition of eukaryotic initiation sites.

Obviously the (semi)conserved sequence revealed by a survey such as this need not correspond to the most favorable context for initiation, since the table includes mRNAs that vary in translational efficiency. Nonetheless, reference to the consensus sequence, *especially the highly conserved -3 position,* can be of help when searching a new mRNA sequence to locate the translational initiation site. It is important to avoid two errors when using this approach:

(a) If inspection of the sequence near the 5'-end of the mRNA were to reveal two AUG triplets that conform approximately equally to the consensus sequence, it would be incorrect to conclude that either AUG is equally likely to be the initiator codon. Because 40S ribosomal subunits most likely scan the 5'-end of the mRNA in a linear fashion (Kozak, Cell 34, 971-978, 1983), *the 5'-proximal AUG triplet is the first to be "inspected."* If the sequence preceding the first AUG triplet conforms closely to the consensus, especially if an A occurs in position -3, *the search ends there.* There are two exceptions to this rule. The first involves a small number of mRNAs in which the reading frame following the first ANN*AUG* sequence is short, terminating upstream from a second AUG codon *to which ribosomes seem to gain access by reinitiating!* The second exception consists of a single example: the mRNA derived from influenza B virus genome segment 6 allows ribosomes to initiate efficiently at the first and the second AUG codons, although the first AUG triplet occurs in a "good" context (ANNAUGA) and is not followed by a terminator codon (Shaw et al., Proc. Natl. Acad. Sci. USA 80, 4879-4883, 1983). I have no explanation for this at present.

(b) An AUG triplet that deviates from the consensus in the crucial -3 position *can nevertheless serve as the initiator codon.* This is evidenced by a few mRNAs in the table (entries 40, 98, 129, 133, 134, 196) and also by experimental manipulation of the sequence flanking the initiator codon (Sherman et al., Cell 20, 215-222, 1980; Kozak, manuscript submitted). As a consequence of initiating at a "weak" AUG codon, however, those rare messenger RNAs are predicted to have two special properties: translation should be inefficient; and ribosomes should initiate not only at the first (weak) AUG *but also at the next AUG that lies downstream.* Such mRNAs should therefore have the potential to direct synthesis of two proteins. This has been shown to occur with a few viral mRNAs (Kozak, 1983, op.cit.) but it has yet to be demonstrated for cellular mRNAs.

The -1 to -5 consensus sequence detected in this survey differs from previously-suggested eukaryotic consensus sequences (Hagenbüchle et al., Cell 13, 551-563, 1978; Baralle and Brownlee, Nature 274, 84-87, 1978; Stiles et al., Cell 25, 277-284, 1981) in both its high frequency of occurrence and its constant position relative to the AUG initiator codon. None of the previously-suggested consensus sequences met those criteria. Until further experiments are carried out, it is premature to speculate about the mechanism by which flanking nucleotides might modulate recognition of the AUG initiator codon; but the temptation is irresistable. Sargan et al. (FEBS Lett., 147, 133-136, 1982) have noted an intriguing complementarity between the sequence CCACC in mRNA and the sequence GGUGG at the base of the 3'-terminal hairpin structure in 18S ribosomal RNA. The possibility of base pairing between mRNA and rRNA thus seems worth exploring. An alternative rationalization for the conserved sequence preceding the initiator codon is that it might base-pair with a complementary sequence just downstream from the AUG codon. The resulting hairpin could help to identify the initiation site. Although some mRNAs (see entries 18, 66, 151) have the potential to form a stable hairpin structure centered about the AUG codon, this is by no means universal. Moreover, comparison of closely-related sequences does not reveal compensatory changes that would preserve the potential hairpin structure.

Bibliography. The numbers used here to identify each mRNA correspond to those in column 1 of the table. Bibliographic data are presented in condensed form: first author, year, journal title, volume, first page. Personal communications are indicated by the letters *pc* after an individual's name. My thanks are extended to those individuals.

001. Noda(1982)Nature 299,793.
002. Noda(1983)Nature 302, 528.
003. Noda(1983)Nature 301, 251.
004. Ricca(1981)JBC 256, 11199.
005. Fornwald(1982)NAR 10, 3861.
006. Hanauer(1983)NAR 11, 3503.
007. Zakut(1982)Nature 298, 857.
008. Nudel(1983)NAR 11, 1759.
009. Sanchez(1983)JMB 163, 533.
010. Sanchez, op.cit.
011. Benyajati(1983)Cell 33,125.
012. Rogers(1983)JBC 258, 8169.
013. Hagenbüchle(1980)Cell 21, 179.
014. Hagenbüchle(1981)Nature 289,643.
015. Ohkubo(1983)PNAS 80, 2196.
016. Bock(1982)NAR 10,8113.
017. Amara(1982)Nature 298,240.
018. Putkey(1983)JBC 258, 11864.
019. Hennighausen(1982)EJB 126, 569.
020. Hobbs(1982)NAR 10, 8079.

021. Blackburn(1982)NAR 10, 2295.
022. Hobbs(1982)NAR 10, 8079;
Yu-Lee(1983)JBC 258, 10794.
023. Rodakis(1982)PNAS 79,3551.
024. Jones(1980)Cell 22, 855.
025. Jones, op.cit.
026. Jones, op.cit.
027. Jones, op.cit.
028. Fiddes(1980)Nature 286, 684.
029. Harris(1982)NAR 10,2177.
030. Vogeli(1981)PNAS 78, 5334.
031. Wiebauer(1982)PNAS 79,7077.
032. Jeltsch(1982)EJB 122,291.
033. Furutani(1983)Nature 301,537.
034. Snyder(1982)Cell 29,1027.
035. Snyder, op.cit.
036. Mizukami(1983)PNAS 80,3958.
037. Nunberg(1980)Cell 19,355.
R. Schimke, *pc.*
038. MacDonald(1982)Biochem. 21,1453.
039. MacDonald, op.cit.
040. Noda(1982)Nature 297,431.
Comb(1982)Nature 295,663.
041. Scott(1983)Science 221,236.
Gray(1983)Nature 303, 722.
042. Gordon(1983)JBC 258,3356.
043. Morinaga(1983)PNAS 80,4604.
044. Law(1981)Nature 291,201.
Eiferman(1981)Nature 294,713.
045. Kant(1983)PNAS 80,3953.
046. Crabtree(1982)Cell 31,159.
047. Tsujimoto(1979)Cell 18,591.
048. Yoo(1982)PNAS 79,1049.
049. Engel(1983)PNAS 80,1392.
050. Erbil(1982)Gene 20,211.
051. Erbil(1983)EMBO 2,1339.
052. Baralle(1977)Cell 12,1085.
Wilson(1980)JBC 255,2807.
053. Proudfoot(1982)Cell 31,553.
054. Baralle(1978)Nature 274,84.
055. Baralle(1977)Nature 267,279.
Heindell(1978)Cell 15,43.
056. Kay(1983)NAR 11,1537.
057. Dolan(1983)JBC 258,3983.
058. Hampe(1981)Gene 14,11.
059. Roninson(1981)PNAS 78,4782.
060. Haynes(1980)PNAS 77,7127.
061. Baralle(1977)Cell 12,1085.
062. Slightom(1980)Cell 21,627.
063. Baralle(1980)Cell 21,621.
064. Konkel(1978)Cell 15,1125.
Baralle(1978)Nature 274,84.
065. Hansen(1982)JBC 257,1048.
066. Baralle(1977)Cell 10,549.
Efstratiadis(1977)Cell 10,571.
067. Hardison(1981)JBC 256,11780.
068. Patient(1983)JBC 258,8521.
069. Banville(1983)JBC 258,7924.
070. Laperche(1983)Cell 32,453.
071. Lund(1982)PNAS 79,345.
072. Lund(1983)JBC 258,3280.
073. Bell(1983)Nature 302,716.
074. Muskavitch(1982)Cell 29,1041.
075. Garfinkel(1983)JMB 168,765.
076. Garfinkel, op,cit.
077. Garfinkel, op.cit.
078. Fiddes(1979)Nature 281,351.
Fiddes(1981)JMAG 1,3.
079. Chin(1981)PNAS 78,5329.
080. Miller(1980)JBC 255,7521.
Woychik(1982)NAR 10,7197.
081. DeNoto(1981)NAR 9,3719.
082. Page(1981)NAR 9,2087.
083. Török(1980)NAR 8,3105.
Ingolia(1980)Cell 21,669.
084-087. Ingolia(1981)NAR 9,1627.
Southgate(1983)JMB 165,35.
088. Lalanne(1983)NAR 11,1567.
Kvist(1983)EMBO 2,245.
089. Mathis(1983)Cell 32,745.
090. Malissen(1983)Science 221,750.
091. Long(1983)EMBO 2,389.
092. Sugarman(1983)JBC 258,9005.
093. D'Andrea(1981)NAR 9,3119.
094. Harvey(1983)PNAS 80,2819.
095. Grandy(1982)JBC 257,8577.
096. Sugarman(1983)JBC 258,9005.
097. Ruiz-Vazquez(1982)NAR 10,2093.
Krieg(1983)NAR 11,619.
098. Clark(1981)NAR 9,1583.
099. Heintz(1981)Cell 24,661.
100-104. Sures(1980)PNAS 77,1265.
105-108. Busslinger(1980)NAR 8,957.
Hentschel(1980)Nature 285,147.
109-110. Childs(1982)Cell 31,383.
111. Turner(1983)NAR 11,4093.
112-113. Moorman(1982)FEBS Lett.144,235.
114-115. Moorman(1981)FEBS Lett.136,45.
116. Jolly(1983)PNAS 80,477.
117. Konecki(1982)NAR 10,6763.
118. Selsing(1981)Cell 25,47.
119. Kelley(1982)Cell 29,681.
120. Bernard(1978)Cell 15,1133.
Picard(1983)PNAS 80,417.
121. Early(1980)Cell 19,981.
122. Kataoka(1982)JBC 257,277.
123. Kenten(1982)PNAS 79,6661.
124. Hellman(1982)NAR 10,6041.
125. Perler(1980)Cell 20,555.
126. Bell(1980)Nature 284,26.
127. Lomedico(1979)Cell 18,545.
Cordell(1979)Cell 18,533.
128. Hobart(1980)Science 210,1360.
129. Chan(1981)JBC 256,7595.

130. Sorokin(1982)Gene 20,367.
131. Goeddel(1980)Nature 287,411.
Lawn(1981)PNAS 78,5435.
132. Nagata(1980)Nature 287,401.
133-134. Goeddel(1981)Nature 290,20.
Lawn(1981)Science 212,1159.
135. Shaw(1983)NAR 11,555.
136. Houghton(1980)NAR 8,1913.
Ohno(1981)PNAS 78,5305.
137. Higashi(1983)JBC 258,9522.
138. Derynck(1982)NAR 10,3605.
Gray(1982)Nature 298,859.
139. Devos(1983)NAR 11,4307.
Taniguchi(1983)Nature 302,305.
140. Mason(1983)Nature 303,300.
141. Swift(1982)PNAS 79,7263.
142. Steinert(1983)Nature 302,794.
143. Powell(1983)NAR 11,5327.
144. Nawa(1983)PNAS 80,90.
145. Hall(1982)NAR 10,3503.
146. Hoffman(1982)NAR 10,7819.
147. Brisson(1982)PNAS 79,4055.
148. van het Schip(1983)NAR 11,2529.
149. Cheung(1983)NAR 11,3703.
150. McLean(1983)JBC 258,8993.
151. Grez(1981)Cell 25,743.
Jung(1980)PNAS 77,5759.
152. Vlasak(1983)EJB 135,123.
153. Karin(1982)Nature 299,797,
154. Glanville(1981)Nature 292,267.
155. Daniel(1983)EMBO 2,1061.
156. Blanchetot(1983)Nature 301,732.
157. Nabeshima(1982)NAR 10,6099.
158. Ullrich(1983)Nature 303,821.
159. VanStraaten(1983)PNAS 80,3183.
160. Watt(1983)Nature 303,725;
PNAS 80,6307.
161. Capon(1983)Nature 304,507.
162. Takeya(1983)Cell 32,881.
163. Nakanishi(1981)EJB 115,429.
164. Whitfeld(1982)DNA 1,133.
Cochet(1982)Nature 297,335.
165. McReynolds(1978)Nature 273,723.
166. Catterall(1980)J.Cell Biol.87,480.
167. Kronenberg(1979)PNAS 76,4981.
Weaver(1982)Mol.Cell.Endocrin.
28,411.
168. Hendy(1981)PNAS 78,7365.
Vasicek(1983)PNAS 80,2127.
169. Sogawa(1983)JBC 258,5306.
170. Slightom(1983)PNAS 80,1897.
171. Michelson(1983)PNAS 80,472.
172. Barrera-Saldaña(1983)JBC 258,3787.
173. Pennica(1983)Nature 301,214.
174. Cooke(1980)JBC 255,6502.
Cooke(1982)Nature 297,603.
175. Sasavage(1982)JBC 257,678.
176. Gregory(1982)NAR 10,7581.
177. Lonberg(1983)PNAS 80,3661.
178. Hudson(1981)Nature 291,127.
179. Panthier(1982)Nature 298,90.
180. MacDonald(1982)JBC 257,14582.
181. Amaldi(1982)Gene 17,311.
182. Berry-Lowe(1982)JMAG 1,483.
183. Kandala(1983)NAR 11,3169.
184. Okamoto(1982)JBC 257,15192.
185. Haché(1983)JBC 258,4556.
186. Dugaiczyk(1982)PNAS 79,71.
187. Gubler(1983)PNAS 80,4311.
188. Hobart(1980)Nature 288,137.
189. Magazin(1982)PNAS 79,5152.
190. Minth(1982)JBC 257,10372.
191. Shen(1982)PNAS 79,4575.
192. Funckes(1983)JBC 258,8781.
193. Parker(1982)NAR 10,5121.
194. Parker, op.cit.
195. Hurst(1983)EMBO 2,769.
196. Edens(1982)Gene 18,1.
197. Gurr(1983)PNAS 80,2122.
198. MacDonald(1982)JBC 257,9724.
199. Lemischka(1982)Nature 300,330.
200. Valenzuela(1981)Nature 289,650.
201. Lee(1983)Cell 33,477.
202. Suske(1983)NAR 11,2257.
Chandra(1981)DNA 1,19.
203. Itoh(1983)Nature 304,547.
204. Land(1982)Nature 295,299.
205. Schmale(1983)EMBO 2,763.
206. Geiser(1983)JBC 258,9024.
207. Hung(1981)NAR 9,6407.
208. Hung(1983)JMB 164,481.
209. Hennighausen(1982)NAR 10,2677.
A. Sippel, *pc*.
210. Pedersen(1982)Cell 29,1015.
211. Marks(1982)JBC 257,9976.

DEX-0772

6B

Volume 9 Number 20 1981

Nucleic Acids Research

Possible role of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic ribosomes

Marilyn Kozak

Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA 15260, USA

Received 3 August 1981

ABSTRACT

Sequences flanking the initiator codon in eukaryotic mRNAs are not random. Out of 153 messages examined, 151 have either a purine in position -3, or a G in position +4, or both. Thus, $^{A}_{G}$XXAUGG emerges as the favored sequence for eukaryotic initiation sites. Nucleotides flanking nonfunctional AUG triplets, which occur in the 5'-noncoding region of a few eukaryotic messages, are different from those found at most functional sites. Whereas most authentic initiator codons are preceded by a purine (usually A) in position -3, most nonfunctional AUGs have a pyrimidine in that position. The observed asymmetry suggests that purines in positions -3 and +4 might facilitate recognition of the AUG codon during formation of initiation complexes. To test this idea, in vitro binding studies were carried out with ^{32}P-labeled oligonucleotides. Binding of AUG-containing oligonucleotides to wheat germ ribosomes was significantly enhanced by placing a purine in position -3 or +4. The scanning model, which postulates that 40S ribosomal subunits attach at the 5'-end of a message and migrate down to the AUG codon, is discussed in light of these new observations. A modified version of the scanning mechanism is proposed.

INTRODUCTION

The pivotal role of the AUG codon in defining the start site for protein biosynthesis has long been recognized. Only a small fraction of the AUG triplets in a given message function as ribosome binding sites, however. In prokaryotic messenger RNAs, the major ancillary signal that dictates which AUG codons will be selected by ribosomes is a purine-rich sequence centered about ten nucleotides upstream from the AUG triplet (1). Other sequences located farther to the left (2-4) or right (5,6) of the AUG codon have also been shown to influence translational efficiency, at least in some messages. In prokaryotes, the role of the purine-rich sequence preceding the initiator codon was deduced from comparison of nucleotide sequences among a large number of messages, and from manipulation of messenger RNAs; i.e., altering the "Shine/Dalgarno" region preceding a particular initiator codon lowered or abolished binding of ribosomes to that site (7,8). By contrast, eukaryotic ribosomes appear to be more tolerant of sequence changes within the region of a message preceding the initiator codon (9-12). Comparison of the 5'-proximal nucleotide sequences of eukaryotic messages reveals remarkable heterogeneity, even among some closely related mRNAs (13-15), reinforcing the impression that eukaryotic ribosomes might not require a particular sequence to demarcate the initiation site. Based in part on observations such as these, I have hypothesized that eukaryotic ribosomes bind initially at the 5'-end of a message and then migrate, scanning the mRNA sequence until they encounter the first AUG triplet which, solely by virtue of its position, is the initiator codon. Considerable evidence has been adduced in favor of this "scanning mechanism" (16,17). A compilation in Dec., 1980, revealed that in 90 out of 99 eukaryotic messages which had been sequenced, translation does indeed begin at the AUG triplet which is closest to the 5'-terminus (17). Although it is gratifying that 90% of the messages examined conform to the prediction, the 9 exceptional messages (a number which has since grown to 11) in which translation does *not* start at the first AUG have been puzzling. The scanning model, in its simplest form, states that the initiator codon is recognized by its position (i.e., first-in-line) irrespective of the flanking sequences. But as sequence information has become available from more and more eukaryotic messages, it has become obvious that the nucleotides flanking the initiator codon are not random. The data compiled in this paper show that nucleotides in positions -3 and +4 are highly conserved. (The numbering system used here is $\overset{-3}{X}pXpXp\overset{+1}{A}pUpGp\overset{+4}{X}$.) To determine whether the conserved nucleotides play a role during initiation, I have constructed a series of AUG-containing oligonucleotides and measured their ability to bind to wheat germ ribosomes in vitro. The binding efficiency of the oligonucleotides was significantly enhanced by placing a purine in position -3 or +4. A slightly more elaborate version of the scanning model, which takes these new data into account, may provide an explanation for those exceptional eukaryotic messages in which initiation is not restricted to the 5'-proximal AUG codon.

MATERIALS AND METHODS

Synthesis and characterization of oligonucleotides

To simplify the representation of nucleotide sequences, the designation $\overset{*}{p}$ is used for [^{32}P]; Y = pyrimidine; R = purine; X = any one of the four common ribonucleotides. All of the di- and trinucleotides used in this work were purchased from P-L Biochemicals except for GpUpG, which was from Boehringer Mannheim, and CpCpC, which was from Collaborative Research.

Stepwise synthesis of oligonucleotides varying in position -3

The oligonucleotide ApUpG$\overset{*}{p}$Cp was first synthesized by ligation of [5'-^{32}P] pCp to ApUpG. The donor [5'-^{32}P]pCp was prepared in a reaction containing 40 mM Tris·HCl (pH 8.5), 10 mM dithiothreitol, 10 mM $MgCl_2$, 2 mM spermine, 1 μg bovine serum albumin, 0.2 mM 3'-CMP, 300 μCi of γ-^{32}P-ATP (New England Nuclear, specific activity adjusted to 300 Ci/mmol), and 3 U of polynucleotide kinase (P-L Biochemicals). After incubation at 37°C for 40 min, the reaction was terminated by boiling for 2.5 min. Following inactivation of the polynucleotide kinase, the solution containing [5'-^{32}P]pCp was used immediately in a reaction with RNA ligase. The 40 μl reaction mixture contained 20 μl of (boiled) kinase reaction, 0.6 OD_{260} units of ApUpG triplet, 15 U of RNA ligase (from T4 phage-infected E. coli; P-L Biochemicals product 0880), 0.13 mM ATP, 10% dimethylsulfoxide (Mallinckrodt), additional $MgCl_2$ to give a final concentration of 20 mM, and additional Tris·HCl to give a final concentration of 50 mM. Incubation was carried out at 4°C for 18-20 hr, as recommended by Bruce and Uhlenbeck (18). Unless otherwise indicated in the text, these reaction conditions permitted quantitative ligation of the ^{32}P-labeled donor to the acceptor oligonucleotide. The product of the ligase reaction was purified by phenol extraction followed by electrophoresis on Whatman 3MM paper in pyridine/acetate buffer at pH 3.5. ApUpG$\overset{*}{p}$Cp migrates slightly faster than the xylene cyanol marker. The ^{32}P-labeled oligonucleotide was eluted from the paper with water, further purified by chromatography on Bio-Gel P-2, then stored in water at -70°C.

Sequential kinase/ligase reactions were carried out a second time to obtain XpCpCpApUpG$\overset{*}{p}$Cp, where X = C, A or G. The tetranucleotide ApUpG$\overset{*}{p}$Cp from the preceding step was first phosphorylated by incubation with polynucleotide kinase and 0.2 mM (nonradioactive) ATP; the resulting pApUpG$\overset{*}{p}$Cp was used as donor in a reaction with RNA ligase. Three ligase reactions were carried out, using as acceptor either CpCpC, ApCpC or GpCpC; reaction conditions were as described in the preceding paragraph. After phenol extraction, the ^{32}P-labeled heptanucleotides were recovered by precipitation from 70% ethanol.

The kinase/ligase reactions were repeated a third time to obtain CpCpCp-XpCpCpApUpG$\overset{*}{p}$Cp. The heptanucleotide ApCpCpApUpG$\overset{*}{p}$Cp, GpCpCpApUpG$\overset{*}{p}$Cp or CpCpCp-ApUpG$\overset{*}{p}$Cp was first phosphorylated in a standard kinase reaction with nonradioactive ATP. In the subsequent ligase reaction, pXpCpCpApUpG$\overset{*}{p}$Cp (X = C, A or G) served as the donor, with CpCpC (0.6 OD_{260} units/25 μl reaction) as acceptor.

The size and purity of various oligonucleotides was checked by homochromatography on DEAE thin layer plates at 60°C using homomixture c (19). Autoradiography was carried out at room temperature using Kodak BB-1 film.

Synthesis of oligonucleotides varying in position +4

^{32}P-Labeled pentanucleotides of the form ApApUpGpX (where X = C, A, G or U) were synthesized by ligating $\overset{*}{p}$GpX to ApApU. The 5'-phosphorylated dinucleotide donor was first prepared in a reaction with polynucleotide kinase and γ-^{32}P-ATP. Reaction conditions were as described above for synthesis of $\overset{*}{p}$Cp, except that 3'-CMP was replaced with either GpC, GpA, GpG or GpU, each at a concentration of 10 µg/25 µl reaction. After the standard boiling procedure to inactivate kinase, the solution containing [5'-^{32}P]pGpX was used in a ligase reaction, with ApApU (0.6 units/25 µl reaction) as acceptor. These ligase reactions were incomplete, and therefore at the end of the incubation the ^{32}P-labeled pentanucleotides were separated from unreacted donor by preparative electrophoresis on 3MM paper. The pentanucleotides were eluted with water, and a portion of each sample was used directly for ribosome binding. Another aliquot was used as acceptor in a second ligase step, with nonradioactive pCp as donor. This generated a series of hexanucleotides of the form ApApU$\overset{*}{p}$GpXpCp, where X = C, A, G or U.

The structure of the pentanucleotides was confirmed by a series of enzymatic digestions: P1 nuclease yielded 5'-^{32}P-GMP; T2 RNase yielded 3'-^{32}P-UMP; pancreatic RNase yielded a ^{32}P-labeled product that co-migrated with ApApUp on DEAE paper at pH 3.5. The mobilities of the intact dinucleotides and pentanucleotides on 3MM paper at pH 3.5 (see Fig. 2) are consistent with their composition; that is, the mobilities follow the expected gradient U>G>A>C.

Binding of oligonucleotides to wheat germ ribosomes

The ability of ^{32}P-labeled oligonucleotides to form 80S initiation complexes was measured using conditions that were previously employed with natural messenger RNAs (20). The wheat germ S23 extract was supplemented with 1 mM ATP, 0.24 mM GTP, 200 µM sparsomycin, and other components as described (20). The final concentration of magnesium was 2.8 mM. After incubation for 10 min at 19°C, samples were chilled and layered onto 10-30% glycerol gradients. Centrifugation was for 3 hr, 39,000 rpm, 4°C in the Beckman SW41 rotor. Gradient fractions (0.4 ml) were mixed with 0.4 ml of water and 8 ml of Beckman HP scintillation cocktail, and were counted to determine the distribution of radioactivity.

RESULTS

A survey of nucleotides flanking the initiator codon in eukaryotic mRNAs

If nucleotides bordering the AUG codon are involved in defining the ribosome binding site, comparison of a large number of initiation sites might reveal a conserved sequence. The simplest approach is to examine the frequency of occurrence of a given nucleotide in each position to the left and right of the AUG codon. Figure 1 presents such an analysis for 106 eukaryotic messages. The region selected for analysis encompasses 15 nucleotides preceding the initiator codon, and the first 15 nucleotides of the coding region. This is the portion of the message that would be protected against nuclease by an 80S ribosome bound at the initiator AUG. There are a few remarkable features in the nucleotide distribution shown in Figure 1. Within the *noncoding* region (residues -1 to -15), the greatest disproportion occurs close to the AUG triplet: 80% of eukaryotic messages have an A in position -3, and C residues occur with high frequency in positions -1, -2, -4 and -5. The G content is unusually low throughout region -1 to -15. In the *coding* region depicted in Figure 1, the nucleotide distribution is approximately random except for positions +4 (60% G), +5 (44% C) and +6 (42% U). The significance of these asymmetries is not immediately obvious, nor is the problem easily approached experimentally. To simplify the task, I have focused in this study on the two positions which show the greatest deviation: position -3, which is a purine (most often A) in 94% of the messages examined; and position +4, which is also a purine (most often G) in 83% of eukaryotic messages.

In attempting to assign a function to these conserved nucleotides, one is confronted by the problem that some messages (albeit a minority) lack the "consensus sequence." One solution that comes to mind is that, in order to

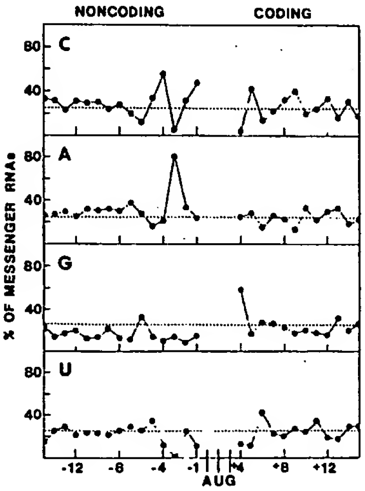


Figure 1. Frequency of occurrence of cytosine (C), adenine (A), guanine (G) and uracil (U) in positions -1 to -15 (preceding the AUG initiator codon), and in coding positions +4 to +15. The AUG triplet is numbered +1 to +3. The dotted line across each panel indicates the 25% value that would be expected on a random basis. For these calculations, I used 90 eukaryotic messages tabulated in reference 17, and 16 additional sequences given in references 21-37. I did not include the human β- and δ-globin sequences, which are very similar to rabbit β-globin; the minor species of mouse β-globin, which is nearly identical to mouse β-globin major; and sequences from the New Jersey strain of vesicular stomatitis virus, which are similar to the Indiana strain. SV40 late 16S mRNA was also omitted because of ambiguity in identifying the initiator codon for VP1.

TABLE 1. GROUPING OF EUKARYOTIC mRNAs ACCORDING TO SEQUENCES FLANKING THE AUG CODON

	Distribution of Functional AUG Initiator Codons[a]		Distribution of nonfunctional AUGs in 5'-Noncoding Region[b]
AXXAUGG	-chicken ovomucoid	-4 actin genes in Dictyostelium (42)	none
	-chicken ovalbumin	-yeast actin, glyceraldehyde dehydrogenase, iso-2-cytochrome c (14), his-4 (43) and enolase (44)	
	-rabbit α- and β-globins	-human leukocyte interferons A B D (41)	
	-mouse α-globin	-human Ig V_H gene segment (46)	
	-mouse β-globin, major & minor	-2 mouse Ig V_H gene segments (21,23)	
	-human α, β, δ, and ε-globins	-mouse Ig $V_{\lambda II}$ gene segment (25)	
	-chicken α-globin (38)	-human chorionic gonadotropin, β (32)	
	-rat preproinsulin I	-chicken VLD lipoprotein II (36)	
	-chicken preproinsulin	-adenovirus-2 hexon protein	
	-anglerfish preproinsulin	-adenovirus-2 protein VI (47)	
	-mouse dihydrofolate reductase	-Rous sarcoma *gag* & *src* (28) proteins	
	-S. purpuratus histones H1 H2B H3	-hepatitis B virus surface antigen (31)	
	-P. miliaris histone H3	-human flu virus proteins NP (48), NS, and HA (subtype H2) (49)	
	-silkmoth chorion gene 10a	-3 late proteins of polyoma virus	
	-rat cytochrome c (45)	-VSV NS and L proteins	
	-satellite tobacco necrosis virus	-VSV (New Jersey strain) N protein	
	-turnip yellow mosaic coat protein	-poliovirus (26,27)	
	-tobacco mosaic virus genome	-rat calcitonin precursor (77)	
	-reovirus proteins σNS and μ1		
	-T antigens of polyoma (39), BK virus (40), and SV40		
AXXAUGA	-chicken conalbumin	-alfalfa mosaic virus coat protein	none
	-rat liver α_{2u} globulin (50)	-one of the large reovirus proteins	
	-silkworm fibroin	-VSV M and G proteins	
	-bovine preproparathyroid hormone	-adenovirus-2 fiber protein	
	-P. miliaris histone H1	-Ad-5 and Ad-12 early region E_1a	
	-human fibroblast interferon	-Ad-2 early region 3, 14.5 K (53) and 16K (54) proteins	
	-mouse α-amylase, salivary & panc.	-Semliki Forest virus capsid (30)	
	-chicken lysozyme	-Sindbis virus capsid protein (29)	
	-human (51) and rat prolactin	-human flu NA: subtypes N2 (55), N1 (56)	
	-yeast iso-1-cytochrome c	-human flu HA: subtypes H3 (57), H1 (58)	
	-silkmoth chorion gene 401a	-fowl plague virus matrix protein	
	-mouse Ig V_H gene segment (22)	-fowl plague virus HA protein (59)	
	-flounder antifreeze peptide (52)		

AXXAUGY	-goat β-globin -chicken β-tubulin -S. purpuratus histones H2A & H4 -P. miliaris histones H2A & H4 -Drosophila 70K heat shock protein -rainbow trout protamine (61) -6 actin genes in Drosophila (62)	-bovine corticotropin/lipotropin (35) -2 histone H2B genes in yeast (60) -2 discoidin I genes in Dictyostelium -2 silkmoth chorion genes, A-family -tobacco mosaic virus coat protein -brome mosaic virus coat protein -VSV N protein (Indiana strain)	none
GXXAUGG	-human γ-globin -chicken β-globin -Xenopus β-globin (65) -human preproinsulin -human and rat growth hormones	-human chorionic gonadotropin α-subunit -reovirus proteins σ2, σ3, μNS, μ2 -Moloney MSV oncogene (63,64) -mouse metallothionein-I (78) -SV40 VP3 (66) and agnogene (66,67)	-SV40 16S and 19S mRNA leaders -poliovirus #6 (ref. 27)
GXXAUGA	-mouse Ig V_K(K2) gene segment (24) -mouse α-fetoprotein (34)	-Ad-2 late polypeptide IX -Ad-2 early region 3, 14K protein(53)	none
GXXAUGY	-rat relaxin (33) -human histone H4 (37)	-chicken type I $α_2$ collagen (68)	-poliovirus #1
YXXAUGG	-human histone H3 (69) -alfalfa mosaic virus RNA-3 (70) -reovirus σ1 protein	-Ad-5 early region E_1b -human leuko. interferons C,F,H (41) -herpes thymidine kinase, SV40 VP2	-poliovirus #5, #6 (ref. 26), and #8 -Rous sarcoma virus #1
YXXAUGA	-SV40 VP1 ?		-poliovirus #4 -mouse α-amylase, liver -mouse α-amylase, salivary -Rous sarcoma virus #2 and #3 -chicken preproinsulin
YXXAUGY	-brome mosaic virus RNA-3 ?		-poliovirus #2, #3 and #7 -$α_2$-collagen #1 and #2 -Semliki Forest virus genome

[a]Messages for which no reference is indicated were included in an earlier compilation (17). The abbreviations used are VSV, vesicular stomatitis virus; MSV, murine sarcoma virus; Ad, adenovirus; HA, hemagglutinin; NA, neuraminidase; Ig, immunoglobulin. [b]The rightmost panel deals with mRNAs in which translation does not begin at the 5'-proximal AUG. Eleven such messages have been identified: SV40 late 16S and 19S mRNAs (74,75); the poliovirus genome (26,27); the genome of Rous sarcoma virus (R. Swanstrom & J.M. Bishop, pers.commun.) and two subgenomic mRNAs of RSV (76; J.M. Bishop, pers.commun.); mouse α-amylase mRNAs from salivary gland and liver (71); chicken $α_2$-collagen (68); chicken preproinsulin (72); and the Semliki Forest virus genome (73). *Nonfunctional* upstream AUGs in these 11 messages are listed on the right; the *functional* initiator codon from each message is included in the left side of the table. Some of these mRNAs have only one AUG preceding the initiator codon; others have 2 (collagen) or 3 (Rous); the poliovirus genome has 8 AUGs preceding the presumptive start site for translation. The upstream AUGs are designated #1 (nearest the 5'-terminus), #2, etc. The poliovirus genome was sequenced independently in two laboratories (26,27). Each group reported a different sequence flanking AUG #6, which is therefore listed twice.

function as an efficient initiation signal, an AUG codon must be flanked by *either* a purine in position -3, or a G in position +4. In other words, at least one of the favored flanking nucleotides is required. To evaluate this idea, I have surveyed and grouped 153 eukaryotic messages on the basis of nucleotides occurring in positions -3 and +4. The data are presented in Table 1. (I attempted to include all eukaryotic messages for which adequate sequence data have been published, and in which the functional initiator codon has been identified. Although sequences bordering the AUG codon have been determined for many messages, in some cases the remainder of the 5'-untranslated sequence is not known. Thus, the number of mRNAs used in compiling Table 1 is larger than the number used in Figure 1. In the case of closely related genes, only one member of the set was chosen for Figure 1, whereas Table 1 is more inclusive.) Of the 153 messages listed in Table 1, 120 have an A in position -3. The largest group (64 messages) has the sequence AXXAUG*G*, but AXXAUGA (32 mRNAs) and AXXAUGY (24 mRNAs) also occur with high frequency at functional initiation sites. The table lists 15 messages in which the initiation site has the sequence GXXAUGG, and another 4 with the sequence GXXAUGA. Thus, in 91% of the messages surveyed (139 out of 153), one finds either an A in position -3 (with G, A or Y in position +4), or one finds a G in position -3 and a purine in position +4. Judging from their high frequency of occurrence, those sequences must function well. The infrequency of finding GXXAUGY or YXXAUGX, on the other hand, might mean that an AUG triplet flanked by those nucleotides functions poorly as an initiation signal.[1]

It is intriguing to note that, in the eleven unusual messages in which translation does *not* begin at the *first* AUG, the pattern of nucleotides flanking the nonfunctional upstream AUG triplets (right side of Table 1) is different from that found at most functional initiation sites. The nonfunctional AUG triplets which occur within the 5'-noncoding regions are clustered in the lower right quadrant of Table 1; i.e., they are bordered by nucleotides which are rarely seen around functional initiator codons. The only serious exceptions are SV40 16S and 19S mRNAs, in which the upstream AUG triplets in the leader are flanked by GXXAUGG--a sequence frequently found at functional initiation sites. But the first AUG in SV40 16S mRNA really should not be listed with the "nonfunctional" AUGs in the right side of Table 1. It was recently shown that ribosomes do initiate at that site, translating the so-called agnogene (67). The mechanism by which ribosomes are also able to initiate at a downstream AUG triplet to make the VP1 protein is not understood; but the ideas proposed in this paper are not contradicted by finding "good" sequences flanking the first AUG in SV40 16S mRNA. The upstream AUG in that message *is functional*. The only other entry in the GXXAUGG column is the sixth AUG in the poliovirus genome. As explained in the footnote to Table 1, two different sequences were reported for that site. Until the ambiguity can be cleared up, it seems fair to omit AUG #6 from further discussion. With these caveats, the data presented in Table 1 suggest the generalization that nonfunctional AUG triplets, which are found in the 5'-noncoding region of a few eukaryotic messages, are bordered by sequences (GXXAUGY or YXXAUGX) which differ from those found around most functional initiator codons.

The main conclusion from this survey is that nucleotides flanking functional initiator codons in eukaryotic messages are not random. Purines occur with very high frequency in positions -3 and +4. As a first step in determining whether the conserved flanking nucleotides play a role during initiation, I have carried out in vitro ribosome binding studies using various synthetic oligonucleotides.

Binding of ^{32}P-labeled oligonucleotides to wheat germ ribosomes

Effect of varying the nucleotide in position +4

Oligonucleotides of the form ApApU$\overset{*}{p}$GpX (where X = C, A, G or U) were constructed by ligating $\overset{*}{p}$GpX to the triplet ApApU. The efficiency of ligation was approximately 40% for reactions with $\overset{*}{p}$GpC and $\overset{*}{p}$GpA, and somewhat less than that for reactions with $\overset{*}{p}$GpG and $\overset{*}{p}$GpU, as shown in Figure 2. The ^{32}P-labeled pentanucleotides, purified by electrophoretic fractionation, were incubated with wheat germ ribosomes under conditions that permitted formation of 80S initiation complexes. As shown in Table 2 (experiments 1 and 2), the efficiency of binding varied from 0.5% for ApApU$\overset{*}{p}$GpU, to 7-10% for ApApU$\overset{*}{p}$GpG.

The pentanucleotide series was converted to hexanucleotides by ligating pCp to the 3'-terminus. The chromatographic analysis shown in Figure 3 reveals that the ligation was quantitative: all ^{32}P-radioactivity was converted to the slower-migrating hexanucleotide position. The overall efficiency of binding to ribosomes was higher with the hexanucleotide series than with the pentanucleotides. Table 2 (experiment 3) shows that, within the hexanucleotide series ApApU$\overset{*}{p}$GpXpCp, varying the nucleotide in position +4 produced a gradient in binding efficiency: G>A>C>U. A similar gradient was *not* observed with the control series ApApU$\overset{*}{p}$GpUpXp (X = C, A, G or U; experiment 4). Thus, the efficiency of binding was markedly enhanced by placing a purine in position +4, but not in position +5.

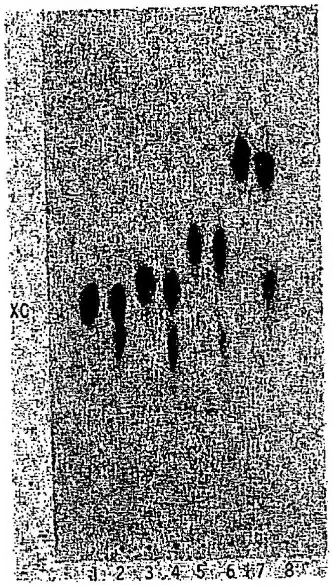


Figure 2. Synthesis of $\mathrm{ApApU\overset{*}{p}GpX}$ series of oligonucleotides. Odd-numbered lanes show the ^{32}P-labeled dinucleotides formed in the first-step kinase reactions: lane 1, $\mathrm{\overset{*}{p}GpC}$; lane 3, $\mathrm{\overset{*}{p}GpA}$; lane 5, $\mathrm{\overset{*}{p}GpG}$; lane 7, $\mathrm{\overset{*}{p}GpU}$. The even-numbered lanes show the products of the second-step ligase reactions, with nonradioactive ApApU as acceptor: lane 2, $\mathrm{ApApU\overset{*}{p}GpC}$; lane 4, $\mathrm{ApApU\overset{*}{p}GpA}$; lane 6, $\mathrm{ApApU\overset{*}{p}GpG}$; lane 8, $\mathrm{ApApU\overset{*}{p}GpU}$. The pentanucleotide is the slower spot in lanes 2, 4, 6 and 8; the faster spot in each lane represents residual unligated dinucleotide. Products were fractionated by electrophoresis on Whatman 3MM paper at pH 3.5. An autoradiogram is shown. XC - xylene cyanol marker. The origin is at the bottom.

TABLE 2. Binding of ^{32}P-labeled oligonucleotides to wheat germ 80S ribosomes: effect of varying the nucleotide in position +4 or +5

Experiment	^{32}P-labeled pentanucleotide	Percent bound[a]	Experiment	^{32}P-labeled hexanucleotide	Percent bound[a]
1	$\mathrm{ApApU\overset{*}{p}Gp\underline{C}}$	3	3	$\mathrm{ApApU\overset{*}{p}Gp\underline{C}pCp}$	12
	$\mathrm{ApApU\overset{*}{p}Gp\underline{A}}$	4		$\mathrm{ApApU\overset{*}{p}Gp\underline{A}pCp}$	18
	$\mathrm{ApApU\overset{*}{p}Gp\underline{G}}$	7		$\mathrm{ApApU\overset{*}{p}Gp\underline{G}pCp}$	23
	$\mathrm{ApApU\overset{*}{p}Gp\underline{U}}$	0.5		$\mathrm{ApApU\overset{*}{p}Gp\underline{U}pCp}$	6
2	$\mathrm{ApApU\overset{*}{p}Gp\underline{U}}$	0.5	4	$\mathrm{Ap\underline{G}pU\overset{*}{p}GpGpCp}$ (control)	0.3
	$\mathrm{ApApU\overset{*}{p}Gp\underline{G}}$	9.5		$\mathrm{ApApU\overset{*}{p}GpGpCp}$	24
	$\mathrm{Ap\underline{G}pU\overset{*}{p}GpG}$ (control)	0		$\mathrm{ApApU\overset{*}{p}GpUp\underline{C}p}$	5
				$\mathrm{ApApU\overset{*}{p}GpUp\underline{A}p}$	1
				$\mathrm{ApApU\overset{*}{p}GpUp\underline{G}p}$	3
				$\mathrm{ApApU\overset{*}{p}GpUp\underline{U}p}$	1

[a] After incubation for 10 min at 19°C, each 50 µl reaction mixture was centrifuged through a glycerol gradient. The ^{32}P-radioactivity co-sedimenting with 80S ribosomes is shown as a percent of the total radioactivity recovered. Each sample contained 30,000 cpm of ^{32}P-oligonucleotide.

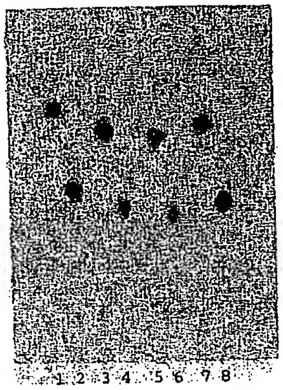


Figure 3. Addition of pCp to the 3'-end of the pentanucleotide series ApApU$\overset{*}{p}$GpX. The ^{32}P-labeled pentanucleotides ApApU$\overset{*}{p}$GpC, ApApU$\overset{*}{p}$GpA, ApApU$\overset{*}{p}$GpG and ApApU$\overset{*}{p}$GpU are shown in lanes 1, 3, 5 and 7, respectively. The products obtained after ligation to (nonradioactive) pCp are shown in lanes 2, 4, 6 and 8. Fractionation was by homochromatography on a DEAE-cellulose thin layer plate. The origin is at the bottom.

Effect of varying the nucleotide in position -3

The range of oligonucleotides that could be constructed for this study was limited by availability of the required trinucleotide precursors and, more importantly, by the specificity of RNA ligase. As reported previously (79,80), the enzyme has marked sequence preferences with respect to both donor and acceptor molecules. Molecules with 5'-terminal cytidine are, by far, the most efficient donors in reactions catalyzed by T4 RNA ligase. The best trinucleotide acceptor is ApApA; ApCpC functions adequately as acceptor if the enzyme and substrate concentrations are high; ApUpU and UpUpU were very inefficient in preliminary experiments, and therefore they were abandoned. Acceptor activity depends not only on the 3'-terminal residue, but also on the adjacent nucleotides.

To test the idea that a purine in position -3 might enhance binding of AUG-containing oligonucleotides, I first constructed a series of heptanucleotides of the form XpCpCpApUpG$\overset{*}{p}$Cp, where X = C, A or G. These were prepared by ligating ^{32}P-labeled pApUpG$\overset{*}{p}$Cp to either CpCpC, ApCpC or GpCpC. Figure 4 (lanes 2-4) shows that all of the ^{32}P-labeled donor was converted to slower-migrating heptanucleotides. When these oligonucleotides were tested for ability to bind to wheat germ ribosomes, ApCpCpApUpG$\overset{*}{p}$Cp and GpCpCpApUpG$\overset{*}{p}$Cp were slightly more efficient than CpCpCpApUpG$\overset{*}{p}$Cp, as shown in Table 3.

It seemed possible that the stabilizing effect of a purine in position -3 might be more obvious if the purine were not right at the end of the oligonucleotide. To obtain longer templates, I ligated pXpCpCpApUpG$\overset{*}{p}$Cp (X = C, A or G) to CpCpC. The reactions with pCpCpCpApUpG$\overset{*}{p}$Cp and pApCpCpApUpG$\overset{*}{p}$Cp proceeded to

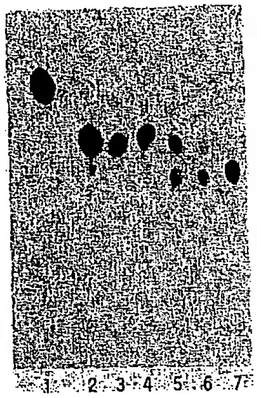


Figure 4. Sequential joining of ^{32}P-labeled pApUpG$\overset{*}{p}$Cp to XpCpC (X = G, A or C), and then to CpCpC. Lane 1, ApUpG$\overset{*}{p}$C; lane 2, GpCpCp-ApUpG$\overset{*}{p}$C; lane 3, ApCpCpApUpG$\overset{*}{p}$C; lane 4, CpCpCp-ApUpG$\overset{*}{p}$C; lane 5, CpCpCpGpCpCpApUpG$\overset{*}{p}$C (slower spot) and residual GpCpCpApUpG$\overset{*}{p}$C; lane 6, CpCpCpApCpCpApUpG$\overset{*}{p}$C; lane 7, (Cp)$_6$ApUpG$\overset{*}{p}$C. The reactions catalyzed by RNA ligase are described in Materials and Methods. Fractionation was by homochromatography on a DEAE-cellulose thin layer plate. The origin is at the bottom. To enhance the resolution, terminal phosphates were removed by incubating all oligonucleotides with bacterial alkaline phosphatase prior to chromatography.

completion, as shown in Figure 4, lanes 6 and 7. With pGpCpCpApUpG$\overset{*}{p}$Cp as donor, however, the extent of joining was only 40% after the first ligation (Figure 4, lane 5). The reaction with pGpCpCpApUpG$\overset{*}{p}$Cp was complete after a second incubation with RNA ligase (data not shown). The structure of the oligonucleotides (Cp)$_6$ApUpG$\overset{*}{p}$Cp and (Cp)$_3$ ApCpCpApUpG$\overset{*}{p}$Cp was confirmed by analyzing the partial T2 RNase digestion products (Figure 5). Ribosome binding experiments were carried out with (Cp)$_6$ ApUpG$\overset{*}{p}$Cp, (Cp)$_3$ApCpCpApUpG$\overset{*}{p}$Cp and (Cp)$_3$GpCpCpApUpG$\overset{*}{p}$Cp. As shown in Table 3, the extent of binding was 7 to 9-fold higher for oligonucleotides with a purine in position -3. Although this effect was readily demon-

TABLE 3. Binding of ^{32}P-labeled oligonucleotides to wheat germ 80S ribosomes: effect of varying the nucleotide in position -3

^{32}P-labeled oligonucleotide	Percent of input oligonucleotide bound[a]		
	Experiment 1	Experiment 2	Experiment 3
ApUpG$\overset{*}{p}$Cp	3	n.d.	n.d.
$\underline{C}$pCpCpApUpG$\overset{*}{p}$Cp	3	3	3
$\underline{A}$pCpCpApUpG$\overset{*}{p}$Cp	5	5	5
$\underline{G}$pCpCpApUpG$\overset{*}{p}$Cp	6	6	n.d.
ApCpCpApUp$\underline{U}\overset{*}{p}$Cp			0.2
ApCpCp$\underline{G}$pUpG$\overset{*}{p}$Cp			0.2
CpCpCp$\underline{C}$pCpCpApUpG$\overset{*}{p}$Cp		1	1
CpCpCp$\underline{A}$pCpCpApUpG$\overset{*}{p}$Cp		9	7
CpCpCp$\underline{G}$pCpCpApUpG$\overset{*}{p}$Cp		n.d.	9

[a] The experiments were carried out as described in the footnote to Table 2. Three independent preparations of oligonucleotides were tested in experiments 1, 2 and 3. n.d. - not done.

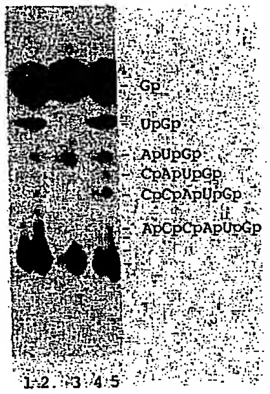


Figure 5. Autoradiogram of products derived by partial hydrolysis of $(Cp)_6ApUpG\overset{*}{p}$ (lanes 1,2) and $(Cp)_3\,ApCpCpApUpG\overset{*}{p}$ (lanes 4,5). Samples were incubated with T2 RNase (5 U/ml) for 15 min (lanes 1,4) or 45 min (lanes 2,5). Partial digestion products were fractionated by homochromatography on DEAE-cellulose. The uniform spacing between spots 3 through 9 in lane 2 is as expected for $(Cp)_6ApUpG\overset{*}{p}$. In lanes 4 and 5, the larger shift between spot 5 ($CpCpApUpG\overset{*}{p}$) and spot 6 ($ApCpCpApUpG\overset{*}{p}$) is consistent with the proposed sequence. [In a homologous partial digestion series, the distance between two oligonucleotides which differ by a purine nucleotide is always larger than that between two oligonucleotides which differ by a pyrimidine nucleotide (81).] The markers in lane 3 are $G\overset{*}{p}$, $ApUpG\overset{*}{p}$ and $(Cp)_6ApUpG\overset{*}{p}$. Prior to carrying out the analysis shown in this figure, the oligonucleotides were digested with T_1 RNase to remove 3'-terminal Cp, leaving ^{32}P directly at the 3'-end of the T_1-derived product. This simplified identification of partial digestion products subsequently obtained with T2 RNase.

strated using the small oligonucleotides described above, with longer templates, such as $(Cp)_{12}$ ApUpGpCp, it was more difficult to show dependence on a purine in position -3. The problem is that longer oligonucleotides bound with considerably higher efficiency (25 to 45%) *seemingly due to just their length.* Not surprisingly, it is difficult to demonstrate a stabilizing effect due to changing one nucleotide when binding has already been dramatically enhanced by the length-effect.

The controls listed in Tables 2 and 3 indicate that, under the conditions of these experiments, only AUG-containing oligonucleotides were able to bind to ribosomes. $Ap\underline{GpU\overset{*}{p}G}pG$ and $Ap\underline{GpU\overset{*}{p}G}pGpCp$ (Table 2), as well as $ApCpCp\underline{GpUpG}\overset{*}{p}Cp$ and $ApCpCp\underline{ApUpU}\overset{*}{p}Cp$ (Table 3) showed negligible binding.

DISCUSSION

A survey of sequences flanking the initiator codon in eukaryotic messenger RNAs reveals that almost all functional AUG triplets are preceded by a purine (usually A) in position -3. A high proportion of functional initiation sites also have a purine (usually G) following the AUG codon. From the data summarized in Table 1, $\genfrac{}{}{0pt}{}{A}{G}$XXAUGG emerges as the favored sequence for eukaryotic initiation sites. Of the 153 messages included in the survey, only 11 have a pyrimidine in position -3, and in 9 of those the AUG codon is followed by G. Thus, only 2 putative initiation sites[2]--those of SV40 VP1 and brome mosaic virus

RNA-3--lack *both* a purine in position -3 and a G residue in position +4. The sequence $^{A}_{G}$XXAUGG which characterizes most functional initiation sites is *never* observed among the nonfunctional AUGs found in the 5'-noncoding region of eukaryotic messages. It is obvious from the data in Table 1, however, that differences in sequences flanking the AUG triplet do not categorically distinguish functional from nonfunctional sites. Although most functional initiator codons are preceded by a purine in position -3 and most nonfunctional AUGs have a pyrimidine in that position, the sequences GXXAUGY and YXXAUGG occur at a small number of functional sites as well as at (presumably) nonfunctional AUGs within the leader region of poliovirus and Rous sarcoma virus RNAs. A mechanism of initiation compatible with this nonunique distribution is outlined below. The main conclusions from the survey presented in Table 1 are (a) that nucleotides flanking *functional* initiator codons (particularly in positions -3 and +4) are not random; and (b) that nucleotides flanking *nonfunctional* AUG triplets which occur within the 5'-untranslated region of a few eukaryotic messages are different from those bordering most functional initiator codons.

This asymmetry suggests that purines in positions -3 and +4 might facilitate recognition of the AUG codon during formation of initiation complexes. The idea gains support from the oligonucleotide binding studies described above. The extent of binding to wheat germ ribosomes was increased several-fold by placing a purine, rather than a pyrimidine, in position +4 (Table 2). The facilitating effect of a purine in position -3 was also readily demonstrated (Table 3), particularly with the series CpCpCpXpCpCpApUpG$\overset{*}{p}$Cp (X = C, A or G). Since only a small number of permutations were tested in this study, it might be that nucleotides in positions other than those tested also contribute to the stability of initiation complexes. But it is encouraging that differences in oligonucleotide binding can be detected upon varying the component in position -3 or +4; a purine seems to be preferred in both positions.

These results cannot be interpreted in isolation. Although the data in this paper suggest that recognition of the initiator codon is influenced by the flanking sequences, the *position* of the AUG triplet (i.e., near the 5'-end of the message) still seems to be the primary determinant of a functional initiation site (16,17). Within the interior of eukaryotic messenger RNAs the sequence $^{A}_{G}$XXAUGG occurs many times; ribosomes do not initiate at these internal AUGs despite the favorable flanking sequence. In apparent contradiction to the studies described above, some previous experiments seemed to indicate that eukaryotic ribosomes select the AUG initiator codon *without regard to the flanking sequences*. Sherman and colleagues (9) showed, for example, that when the normal initiator codon in the yeast cytochrome c gene was inactivated by mutation, introduction of a new AUG triplet almost anywhere within a 37-nucleotide region restored translation. The sequences flanking the ectopic AUG initiator codons in the pseudorevertants varied widely, and often did not correspond to the optimal sequences defined above. How can those experiments be reconciled with the new data described herein? The interpretation I favor is a modified version of the scanning mechanism; namely, that flanking sequences (nucleotides -3 and +4) *modulate the efficiency* with which the migrating 40S ribosomal subunit recognizes an AUG triplet as a "stop signal." *Some* 40S subunits will stop at the first AUG *irrespective of the flanking sequences;* if the nucleotides bordering an AUG codon are *optimal*, virtually *all* 40S subunits will stop at that AUG. Sherman's data are compatible with such a mechanism: *some* cytochrome c is made in the pseudorevertants, indicating (only) that *some* 40S subunits stop at the first AUG, even when it is flanked by suboptimal sequences. Those experiments do not indicate that the ectopic AUG triplets in the pseudorevertants function as efficiently as the wild-type sequence AUAAUGA presumably does. Thus, the genetic experiments are not incompatible with a modified scanning mechanism in which flanking sequences[3] affect the efficiency with which an AUG codon is recognized as a "stop signal."

The proposed mechanism has some interesting implications:

(a) The scanning model in its simplest form (see Introduction) predicts that spurious AUG triplets cannot occur in the region preceding the functional initiation site; recognition of the authentic initiator codon depends on its being first-in-line. This prediction is upheld by most, but not all, eukaryotic messages. The modified scanning model, on the other hand, admits that ribosomes can initiate at a downstream site, provided that all of the upstream AUGs are flanked by "unfavorable" sequences such that *some* 40S ribosomes can get through. The data in the right side of Table 1 provide encouragement for this idea. In those messages in which translation does not begin at the first AUG, almost all of the AUGs which are bypassed have a pyrimidine in position -3.

(b) Although upstream AUGs are not an absolute barrier to initiating downstream, it seems reasonable to expect that occurrence of AUG triplets within the 5'-noncoding region of a message would reduce translational efficiency, since *some* 40S ribosomes would stop at each upstream AUG irrespective of the flanking sequences. Consistent with this idea, poliovirus RNA is a notoriously inefficient message in vitro (84), as is the Rous sarcoma virus genome (85). The notion that upstream AUGs impair translational efficiency would explain why 90% of eukaryotic mRNAs have *no* AUGs in the 5'-noncoding region. Even if the

flanking sequences are such that some ribosomes can get through, upstream AUGs probably have a deleterious effect. Parenthetically, the idea that some 40S subunits stop and initiate at each upstream AUG rationalizes the finding that, in those few messages which have AUG triplets in the "5'-noncoding region" (the semantic difficulty is obvious), the upstream AUGs are almost always followed closely by in-phase terminator codons (17). Thus, ribosomes which initiate prematurely at an upstream AUG are returned quickly to circulation.

(c) The apparent inability of eukaryotic ribosomes to bind to sites in the interior of a message generally means that a eukaryotic mRNA can direct synthesis of only one protein--that encoded nearest the 5'-end of the template. But a single message might direct synthesis of two proteins if the AUG triplet at the start of the first coding region were flanked by unfavorable sequences, such that only some 40S subunits stop and initiate at that site while others advance to the next AUG. Two examples come to mind: SV40 late 19S mRNA is believed to direct synthesis of both VP2 and VP3 (75, 86); and the mRNA encoding herpes thymidine kinase has been shown to direct synthesis of a second smaller protein (87). In both messages the upstream initiator codon, which seems to be "leaky," is preceded by a pyrimidine in position -3: the initiation site for SV40 VP2 is UCCAUGG, and the putative initiation site for herpes thymidine kinase is CGUAUGG.

In summary, it seems likely that nucleotides in positions -3 and +4 influence recognition of the AUG codon by eukaryotic ribosomes, or one of the ribosome-associated components involved in initiation. Binding of synthetic oligonucleotides to wheat germ ribosomes was enhanced 5- to 15-fold by placing a purine in either of those positions. This preference mirrors the observed frequency of nucleotides flanking the initiator codon in natural mRNAs. Although there is no direct evidence concerning the mechanism by which nucleotides bordering the AUG codon facilitate initiation, the mechanism must be compatible with a large body of evidence which suggests that ribosomes attach to eukaryotic mRNAs at an upstream site, and migrate down to the AUG. I have therefore proposed a modified version of the scanning model which postulates that flanking nucleotides modulate the efficiency with which an AUG triplet is recognized as a stop-signal by the migrating 40S subunit. The modified scanning mechanism accounts for those few messages (eleven, to date) in which translation does not begin at the AUG triplet closest to the 5'-terminus, and also for rare messages which seem to direct synthesis of two independently-initiated proteins.

ACKNOWLEDGMENTS

I am grateful to Drs. J.M. Bishop, R. Swanstrom, B. deCrombrugghe and V. Racaniello for providing preprints of their unpublished sequences. The ubiquitous occurrence of adenosine 3 nucleotides before the initiator codon was first pointed out to me by Dr. David Baltimore. This work was supported by grant AI 16634 and Career Development Award AI 00380 from the National Institutes of Health.

FOOTNOTES

[1] The suggested cut-off between efficient and inefficient initiation signals has been set rather arbitrarily between GXXAUGA (which I have called efficient) and GXXAUGY (which I have called inefficient). This division obviously does not follow from the number of functional initiation sites which have those sequences. Both sequences occur infrequently, and therefore both might be viewed as likely-to-be-inefficient. On the other hand, there is no *evidence* that the 4 messages initiating at GXXAUGA or the 3 messages initiating at GXXAUGY (see Table 1) are defective; thus, one might argue that--despite their infrequent occurrence--both GXXAUGA and GXXAUGY should be regarded as efficient. To some extent, I have been guided by the oligonucleotide binding studies which show a purine in position +4 to be much better than a pyrimidine in that site.

[2] There is some uncertainty in pinpointing the initiator codon in these 2 messages. In the case of brome mosaic virus RNA-3, the putative initiation site was identified by the open reading frame that follows; but only a limited portion of that RNA has been sequenced (82). The difficulty in identifying the initiator codon for SV40 VP1 has been discussed previously (17). It is not known whether translation of VP1 in vivo begins at the first or second AUG within the sequence CUUAUGAAGAUGGCC.

[3] Although this paper emphasizes the role of the flanking *primary* sequence in modulating recognition of the AUG codon, other experiments (83) suggest that the secondary/tertiary structure of eukaryotic messages also contributes to the fidelity of initiation. With *denatured* reovirus mRNA as template, 40S ribosomes tend to migrate beyond the first AUG codon--despite the favorable flanking sequences.

REFERENCES

1. Steitz, J.A. (1979) in Biological Regulation and Development, Goldberger, R.F., Ed., pp. 349-399, Plenum Press, New York.
2. Borisova, G.P., Volkova, T.M., Berzin, V., Rosenthal, G. and Gren, E.J. (1979) Nucleic Acids Res. 6, 1761-1774.
3. Cannistraro, V.J. and Kennell, D. (1979) Nature 277, 407-409.
4. Fiil, N.P., Friesen, J.D., Downing, W.L. and Dennis, P.P. (1980) Cell 19, 837-844.
5. Taniguchi, T. and Weissmann, C. (1978) J. Mol. Biol. 118, 533-565.
6. Atkins, J.F., Steitz, J.A., Anderson, C.W. and Model, P. (1979) Cell 18, 247-256.
7. Dunn, J.J., Buzash-Pollert, E. and Studier, F.W. (1978) Proc. Natl. Acad. Sci. USA 75, 2741-2745.
8. Schwartz, M., Roa, M. and Debarbouille, M. (1981) Proc. Natl. Acad. Sci. USA 78, 2937-2941.
9. Sherman, F., Stewart, J.W. and Schweingruber, A.M. (1980) Cell 20, 215-222.
10. Ghosh, P.K., Lebowitz, P., Frisque, R.J. and Gluzman, Y. (1981) Proc. Natl. Acad. Sci. USA 78, 100-104.
11. Barkan, A. and Mertz, J.E. (1981) J. Virol. 37, 730-737.
12. Solnick, D. (1981) Cell 24, 135-143.

13. Firtel, R.A., Timm, R., Kimmel, A.R. and McKeown, M. (1979) Proc. Natl. Acad. Sci. USA 76, 6206-6210.
14. Montgomery, D.L., Leung, D.W., Smith, M., Shalit, P., Faye, G. and Hall, B.D. (1980) Proc. Natl. Acad. Sci. USA 77, 541-545.
15. Jones, C.W. and Kafatos, F.C. (1980) Cell 22, 855-867.
16. Kozak, M. (1980) Cell 22, 7-8.
17. Kozak, M. (1981) in Current Topics in Microbiology and Immunology, Shatkin, A.J., Ed., Vol. 93, pp. 81-123, Springer-Verlag, Berlin.
18. Bruce, A.G. and Uhlenbeck, O.C. (1978) Nucleic Acids Res. 5, 3665-3677.
19. Barrell, B.G. (1971) in Procedures in Nucleic Acid Research, Cantoni, G.L. and Davies, D.R., Eds. Vol.2, pp. 751-779.
20. Kozak, M. and Shatkin, A.J. (1976) J. Biol. Chem. 251, 4259-4266.
21. Sakano, H., Maki, R., Kurosawa, Y., Roeder, W., and Tonegawa, S. (1980) Nature 286, 676-683.
22. Early, P., Huang, H., Davis, M., Calame, K. and Hood, L. (1980) Cell 19, 981-992.
23. Bothwell, A.L.M., Paskind, M., Reth, M., Imanishi-Kari, T., Rajewsky, K. and Baltimore, D. (1981) Cell 24, 625-637.
24. Nishioka, Y. and Leder, P. (1980) J. Biol. Chem. 255, 3691-3694.
25. Tonegawa, S., Maxam, A.M., Tizard, R., Bernard, O. and Gilbert, W. (1978) Proc. Natl. Acad. Sci. USA 75, 1485-1489.
26. Kitamura, N., Semler, B.L., Rothberg, P.G., Larsen, G.R., Adler, C.J., Dorner, A.J., Emini, E.A., Hanecak, R., Lee, J.J., van der Werf, S., Anderson, C.W. and Wimmer, E. (1981) Nature 291, 547-553.
27. Racaniello, V.R. and Baltimore, D. (1981) Proc. Natl. Acad. Sci. USA, in press.
28. Czernilofsky, A.P., Levinson, A.D., Varmus, H.E., Bishop, J.M., Tischer, E. and Goodman, H.M. (1980) Nature 287, 198-203.
29. Rice, C.M. and Strauss, J.H. (1981) Proc. Natl. Acad. Sci. USA 78, 2062-2066.
30. Garoff, H., Frischauf, A.M., Simons, K., Lehrach, H. and Delius, H. (1980) Proc. Natl. Acad. Sci. USA 77, 6376-6380.
31. Valenzuela, P., Gray, P., Quiroga, M., Zaldivar, J., Goodman, H.M. and Rutter, W.J. (1979) Nature 280, 815-819.
32. Fiddes, J.C. and Goodman, H.M. (1980) Nature 286, 684-687.
33. Hudson, P., Haley, J., Cronk, M., Shine, J. and Niall, H. (1981) Nature 291, 127-131.
34. Law, S.W. and Dugaiczyk, A. (1981) Nature 291, 201-205.
35. Nakanishi, S., Teranishi, Y., Watanabe, Y., Notake, M., Noda, M., Kakidani, H., Jingami, H. and Numa, S. (1981) Eur. J. Biochem. 115, 429-438.
36. Wieringa, B., Geert, A.B. and Gruber, M. (1981) Nucleic Acids Res. 9, 489-501.
37. Heintz, N., Zernik, M. and Roeder, R.G. (1981) Cell 24, 661-668.
38. Richards, R.I. and Wells, J.R.E. (1980) J. Biol. Chem. 255, 9306-9311.
39. Friedmann, T., LaPorte, P. and Esty, A. (1978) J. Biol. Chem. 253, 6561-6567.
40. Yang, R.C.A. and Wu, R. (1979) Virology 92, 340-352.
41. Goeddel, D.V., Leung, D.W., Dull, T.J., Gross, M., Lawn, R.M., McCandliss, R., Seeburg, P.H., Ullrich, A., Yelverton, E. and Gray, P.W. (1981) Nature 290, 20-26.
42. McKeown, M. and Firtel, R.A. (1981) Cell 24, 799-807.
43. Farabaugh, P.J. and Fink, G.R. (1980) Nature 286, 352-356.
44. Holland, M.J., Holland, J.P., Thill, G.P. and Jackson, K.A. (1981) J. Biol. Chem. 256, 1385-1395.
45. Scarpulla, R.C., Agne, K.M. and Wu, R. (1981) J. Biol. Chem. 256, 6480-6486.

46. Matthyssens, G. and Rabbitts, T.H. (1980) Proc. Natl. Acad. Sci. USA 77, 6561- 6565.
47. Akusjarvi, G. and Persson, H. (1981) J. Virol. 38, 469-482.
48. Van Rompuy, L., Min Jou, W., Huylebroeck, D., Devos, R. and Fiers, W. (1981) Eur. J. Biochem. 116, 347-353.
49. Air, G.M. (1979) Virology 97, 468-472.
50. Drickamer, K., Kwoh, T.J. and Kurtz, D.T. (1981) J. Biol. Chem. 256, 3634-3636.
51. Cooke, N.E., Coit, D., Shine, J., Baxter, J.D. and Martial, J.A. (1981) J. Biol. Chem. 256, 4007-4016.
52. Lin, Y. and Gross, J.K. (1981) Proc. Natl. Acad. Sci. USA 78, 2825-2829.
53. Hérissé, J. and Galibert, F. (1981) Nucleic Acids Res. 9, 1229-1240.
54. Hérissé, J., Courtois, G. and Galibert, F. (1980) Nucleic Acids Res. 8, 2173-2192.
55. Blok, J. and Air, G.M. (1980) Virology 107, 50-60.
56. Fields, S., Winter, G. and Brownlee, G.G. (1981) Nature 290, 213-217.
57. Min Jou, W., Verhoeyen, M., Devos, R., Saman, E., Fang, R., Huylebroeck, D., Fiers, W., Threlfall, G., Barber, C., Carey, N. and Emtage, S. (1980) Cell 19, 683-696.
58. Winter, G., Fields, S. and Brownlee, G.G. (1981) Nature 292, 72-75.
59. Porter, A.G., Barber, C., Carey, N.H., Hallewell, R.A., Threlfall, G. and Emtage, J.S. (1979) Nature 282, 471-477.
60. Wallis, J.W., Hereford, L. and Grunstein, M. (1980) Cell 22, 799-805.
61. Jenkins, J.R. (1979) Nature 279, 809-811.
62. Fyrberg, E.A., Bond, B.J., Hershey, N.D., Mixter, K.S. and Davidson, N. (1981) Cell 24, 107-116.
63. Reddy, E.P., Smith, M.J., Canaani, E., Robbins, K.C., Tronick, S.R., Zain, S. and Aaronson, S.A. (1980) Proc. Natl. Acad. Sci. USA 77, 5234-5238.
64. VanBeveren, C., Galleshaw, J.A., Jonas, V., Berns, A.J.M., Doolittle, R.F., Donoghue, D.J. and Verma, I.M. (1981) Nature 289, 258-262.
65. Williams, J.G., Kay, R.M. and Patient, R.K. (1980) Nucleic Acids Res. 8, 4247-4258.
66. Reddy, V.B., Thimmappaya, B., Dhar, R., Subramanian, K.N., Zain, B.S., Pan, J., Ghosh, P.K., Celma, M.L. and Weissman, S.M. (1978) Science 200, 494-502.
67. Jay, G., Nomura, S., Anderson, C.W. and Khoury, G. (1981) Nature 291, 346-349.
68. Vogeli, G., Ohkubo, H., Sobel, M.E., Yamada, Y., Pastan, I. and deCrombrugghe, B. (1981) Proc. Natl. Acad. Sci. USA, in press.
69. Clark, S.J., Krieg, P.A. and Wells, J.R.E. (1981) Nucleic Acids Res. 9, 1583-1590.
70. Pinck, M., Fritsch, C., Ravelonandro, M., Thivent, C. and Pinck, L. (1981) Nucleic Acids Res. 9, 1087-1100.
71. Hagenbüchle, O., Tosi, M., Schibler, U., Bovey, R., Wellauer, P.K. and Young, R.A. (1981) Nature 289, 643-646.
72. Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. and Dodgson, J. (1980) Cell 20, 555-566.
73. Wengler, G., Wengler, G. and Gross, H.J. (1979) Nature 282, 754-756.
74. Ghosh, P.K., Reddy, V.B., Swinscoe, J., Choudary, P.V., Lebowitz, P. and Weissman, S.M. (1978) J. Biol. Chem. 253, 3643-3647.
75. Ghosh, P.K., Reddy, V.B., Swinscoe, J., Lebowitz, P. and Weissman, S.M. (1978) J. Mol. Biol. 126, 813-846.
76. Cordell, B., Weiss, S.R., Varmus, H.E. and Bishop, J.M. (1978) Cell 15, 79-91.
77. Jacobs, J.W., Goodman, R.H., Chin, W.W., Dee, P.C., Habener, J.F., Bell, N.H. and Potts, J.T., Jr. (1981) Science 213, 457-459.
78. Glanville, N., Durnam, D.M. and Palmiter, R.D. (1981) Nature 292, 267-269.

79. Ohtsuka, E., Nishikawa, S., Fukumoto, R., Tanaka, S., Markham, A.F., Ikehara, M. and Sugiura, M. (1977) Eur. J. Biochem. 81, 285-291.
80. England, T.E. and Uhlenbeck, O.C. (1978) Biochemistry 17, 2069-2076.
81. Silberklang, M., Gillum, A.M. and RajBhandary, U.L. (1977) Nucleic Acids Res. 4, 4091-4108.
82. Ahlquist, P., Dasgupta, R., Shih, D.S., Zimmern, D. and Kaesberg, P. (1979) Nature 281, 277-282.
83. Kozak, M. (1980) Cell 19, 79-90.
84. Shih, D.S., Shih, C.T., Kew, O., Pallansch, M., Rueckert, R. and Kaesberg, P. (1978) Proc. Natl. Acad. Sci. USA 75, 5807-5811.
85. Pawson, T., Martin, G.S. and Smith, A.E. (1976) J. Virol. 19, 950-967.
86. Piatak, M., Ghosh, P.K., Reddy, V.B., Lebowitz, P. and Weissman, S.M. (1979) in Extrachromosomal DNA. ICN-UCLA Symposia on Molecular and Cellular Biology, Cummings, D.J., Borst, P., Dawid, I.B., Weissman, S.M. and Fox, C.F., Eds.,Vol. XV, pp. 199-215, Academic Press, New York.
87. Preston, C.M. and McGeoch, D.J. (1981) J. Virol. 38, 593-605.

DC 8-0172

6C

Volume 15 Number 20 1987 Nucleic Acids Research

An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs

Marilyn Kozak

Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA 15260, USA

Received July 6, 1987; Revised and Accepted September 23, 1987

ABSTRACT

5'-Noncoding sequences have been compiled from 699 vertebrate mRNAs. (GCC)GCC$^{A}_{G}$CCATGG emerges as the consensus sequence for initiation of translation in vertebrates. The most highly conserved position in that motif is the purine in position -3 (three nucleotides upstream from the ATG codon); 97% of vertebrate mRNAs have a purine, most often A, in that position. The periodical occurrence of G (in positions -3, -6, -9) is discussed. Upstream ATG codons occur in fewer than 10% of vertebrate mRNAs-at-large; a notable exception are oncogene transcripts, two-thirds of which have ATG codons preceding the start of the major open reading frame. The leader sequences of most vertebrate mRNAs fall in the size range of 20 to 100 nucleotides. The significance of shorter and longer 5'-noncoding sequences is discussed.

INTRODUCTION

To search for signals that might influence early steps in translation, I have scrutinized the 5'-noncoding sequences of 699 vertebrate mRNAs, which are identified in the Appendix. The survey included all sequences to which I had access in the published literature except those in which the functional initiator codon had not been clearly identified or where it seemed possible that the cloned cDNA sequence fell short of the true initiator codon. To minimize redundancy, I did not enter every available sequence for large multigene families (especially globins, histones and immunoglobulins), but the sequences that were omitted were usually similar to the ones that were entered; two cases where that is not true are described in footnotes *k* and *n*. When a particular gene was sequenced from more than one organism, I entered both sequences if they differed in at least two positions near the ATG codon. Otherwise I entered only one--the one for which more accessory information was available or, arbitrarily, the human sequence. All mRNA sequences are written with T in place of U since nearly all of the sequences were determined by analyzing DNA.

RESULTS AND DISCUSSION

Context

Previous surveys of eukaryotic mRNA sequences (1, 2) revealed that the sequence flanking functional initiator codons is nonrandom: CC$^{A}_{G}$CCATGG was proposed as the consensus sequence for initiation of translation in higher eukaryotes. The present survey confirms and extends that conclusion using a larger and more diversified data base.

Table 1 and Fig. 1 show a distinctive pattern over the 12 nucleotide stretch preceding the ATG initiator codon. The whole region is deficient in T resi-

Table 1. Frequency of A, C, G and T around the translational start site in vertebrate mRNAs.

POSITION:	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	+4
percent A	23	26	25	23	19	23	17	18	25	**61**	27	15	23
percent C	35	35	35	26	39	37	19	39	**53**	2	49	**55**	16
percent G	23	21	22	33	23	20	**44**	23	15	36	13	21	**46**
percent T	19	18	18	18	19	20	20	20	7	1	11	9	15

Data were compiled from the 699 sequences listed in the Appendix. A window of 12 nucleotides preceding the initiator codon is presented, as well as one nucleotide (position +4) following the ATG codon. The most abundant nucleotide in each position is underlined. Values that are >50% or ≥twice the frequency of the next most abundant nucleotide in that position are shown in boldface.

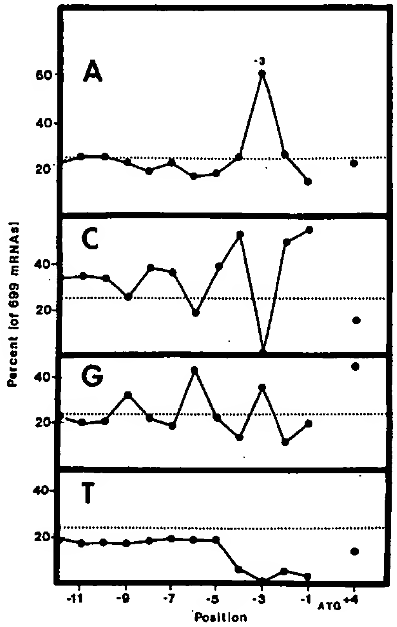


Figure 1. Frequency of A, C, G and T around the ATG initiator codon in 699 vertebrate mRNAs, which are listed in the Appendix. The dotted line across each panel shows the 25% value that would be expected on a random basis.

dues, especially in positions -1 to -4. C is the preferred nucleotide in *every* position except -3, -6 and -9. In positions -6 and -9, G is preferred. Position -3 shows the strongest bias: 61% of vertebrate mRNAs have an A in that position, 36% have G, and only 3% have a pyrimidine. On the 3'-side of the ATG codon G is the preferred flanking nucleotide. Thus the expanded consensus sequence for initiation in vertebrate mRNAs is (GCC)GCC$^{A}_{G}$CC$\underline{\text{ATG}}$G. Site-directed mutagenesis experiments have confirmed the contribution of every nucleotide in positions -1 to -6, as well as the G in position +4 (3, 4), but the significance of the GCC motif in positions -9 to -7 remains to be established. Because mutations in positions -3 and +4 have the strongest influence on translational efficiency, for practical purposes an initiator codon can be designated "strong" or "weak" by considering only those positions. That view is supported by the data in Table 2, in which initiator codons are grouped according to the nucleotides in positions -3 and +4. Among 699 vertebrate mRNAs, only 23 have a pyrimidine 3 nucleotides upstream from the functional initiator codon, and 17 of those "compensate" by having G in position +4. Thus only six mRNAs out of 699 lack the preferred nucleotide in *both* of the crucial positions. One would expect those six mRNAs to be translated very inefficiently, which is not inconsistent with the fact that four of them encode hormones or lymphokines (entries 264, 397, 407 and 650). The other two (entries 172 and 281) are members of large multigene families, for which reason one cannot assess the extent to which the particular mRNA that corresponds to the cloned cDNA is translationally active.

Considerable evidence (3) supports the idea that the ATG codon and flanking sequences function as a "stop signal" for the migrating 40S ribosomal subunit, which binds directly only at the 5'-end of the mRNA. Consequently, *position* as well as context determines which ATG is the functional initiator codon. In cases where there are two ATG codons in equally favorable contexts near the 5'-end of a message, it would be incorrect to conclude that either ATG is equally likely to initiate translation. Theory predicts and experiments confirm (54) that ribosomes initiate exclusively at the 5'-proximal ATG codon when it lies in a favorable context.

The repetition of G in positions -3, -6 and -9 is quite noticeable in

Table 2. Sequences flanking ATG codons in vertebrate mRNAs

Sequence	Functional initiator codons	"Nonfunctional" UPSTREAM ATG codons[a]
-3 +4		
AnnATGG	175	4
A.....A	114	5
A.....C	63	8
A.....T	73	4
G.....G	130	8
G.....A	47	7
G.....C	47	5
G.....T	27	5
C.....G	9	7
C.....A	2	8
C.....Y	4	12
T.....G	8	4
T.....A	0	13
T.....Y	0	16
Total #	699	106 ATGs in 59 mRNAs[b,c]

[a] "Nonfunctional" is a provisional designation. Upstream ATG codons *are* expected to function--an expectation that has been verified with some viral mRNAs but not yet with cellular mRNAs. The more important point is that the indicated upstream ATG codons are not absolute barriers to initiating downstream: a downstream ATG codon starts the major open reading frame in these mRNAs.

[b] The tabulation of upstream ATG codons does not include the 34 oncogenes listed in the Appendix, since they comprise a separate group vis-à-vis the frequency of upstream ATG codons (see text).

[c] Thirty of the upstream ATG codons in this set derive from just four mRNAs: entries 73, 283, 556 and 599. Excepting those four entries and the proto-oncogenes, only 9% of vertebrate mRNAs have upstream ATG codons and they typically have only one.

Table 3. Length distribution of vertebrate mRNA leader sequences.

Length:	<10	10-19	20-29	30-39	40-49	50-59	60-69	70-79	nucleotides
# of mRNAs:	4	10	23	29	36	38	37	40	
Length:	80-89	90-99	100-199	200-299	300-399	400-499	500-599	≥600	nucleotides
# of mRNAs:	15	22	68	19	1	3	2	2	

The table is based on 346 mRNAs for which the transcriptional start site has been mapped. In the case of genes that produce multiple transcripts, only the longest leader was scored. In three cases where ribosomes initiate at the first and second ATG codons (see footnote g in the Appendix), two values were entered in the table; in all three cases there are fewer than 10 nucleotides between the cap and the first functional ATG codon.

Fig. 1. Trifonov (5) has pointed out that there is a strong preference for G in the first position of codons throughout the coding region of both prokaryotic and eukaryotic mRNAs, and he postulates that the periodicity of G residues helps ribosomes stay in frame during translation. An interesting possibility is that "frame monitoring" begins shortly *upstream* from the initiator codon in eukaryotes. In support of that idea, it has been shown by site-directed mutagenesis that correct initiation is strongly favored by placing a purine in position -3 and G in position -6, but the facilitating effect is completely lost when the purines are shifted one nucleotide to the left or right (3, 4).

In recent surveys of mRNA sequences from plants (6), Drosophila (7) and yeast (8), the most striking finding was conservation of a purine--usually A--in position -3. The reported values for position -3 were 53% A and 23% G in 47 plant mRNAs; 82% A and 13% G in 77 Drosophila mRNAs; and 81% A in 96 yeast mRNAs. Although such data encourage the idea that A or G in position -3 somehow favors initiation in all eukaryotic systems, the effect of context on initiation has yet to be tested experimentally in nonvertebrates. The overall A-richness of leader sequences in yeast and plant mRNAs somewhat diminishes the statistical significance of finding A most often in position -3. Leader sequences on Dictyostelium mRNAs are also notoriously A+T rich. This might be a hint that ribosomes from lower eukaryotes and plants are less able to deal with secondary structure than are metazoan ribosomes (9).

Upstream ATG codons

Three points about the occurrence of upstream ATG codons merit comment.

(i) They are relatively rare in vertebrate mRNAs-at-large, as indicated in Table 2. The raw data from which Table 2 was compiled are given in the Appendix.

(ii) The big exception to the foregoing generalization are proto-oncogenes (entries 454 to 487), nearly two-thirds of which produce mRNAs that have ATG codons--usually more than one--preceding the start of the major open reading frame. In view of the inhibitory effect of upstream ATG codons (3), it is probably not an accident that activation of proto-oncogenes by transduction or translocation often deletes the cumbersome leader sequence. Preliminary evidence (10, 11) encourages the hypothesis that the expression of some oncogenes is regulated in part at the level of translation.

(iii) The context around "nonfunctional" upstream ATG codons differs strikingly from the functional initiator codons listed in Table 2. Whereas functional initiator codons are rarely preceded by a pyrimidine in position -3, upstream ATG codons often occur in that unfavorable context. The notion that scanning is "leaky" in such mRNAs--with some 40S ribosomal subunits stopping and initiating at the upstream site while some reach the second ATG codon--is supported by some experimental evidence (3, 12).

Leaky scanning obviously cannot account for the ability of ribosomes to initiate downstream from a strong ATG codon, and a considerable number of the upstream ATG codons in Table 2 do occur in a favorable context. Those ATG codons are nearly always followed by a terminator codon, however, and it seems likely that--after translating the upstream "minicistron" and terminating--ribosomes reinitiate at the next ATG codon downstream (12, 13 and references therein). Given leaky scanning and reinitiation as devices with some experimental justification that permit initiation at an ATG codon that is not "first," *an upstream ATG codon should pose a problem only if it occurs in a favorable context for initiation and is* not *followed by a terminator codon* before the start of the major open reading frame. Among the 699 mRNAs listed in the Appendix, only five have that problematical structure; they are entry 120 (where the upstream ATG codon lies very close to the cap and hence might be inefficient--see below); entries 520, 553 and 599 (where the potentially-inhibitory ATG codon is preceded by a far-upstream minicistron which might have a sparing effect, as explained in reference 13); and entry 247, for which I have no excuse.

It might be noted parenthetically that upstream ATG codons seem to occur more commonly in Drosophila than in vertebrate mRNAs, although I cannot cite precise statistics for Drosophila, nor is it always certain that the ATG-burdened leader sequence belongs to a functional form of mRNA (14).

Leader length

The precise length of the 5'-noncoding sequence is known for about half of the entries in the Appendix, and those mRNAs were used to compile the data in Table 3. Only one-fourth of the mRNAs that were scored have a leader sequence longer than 100 nucleotides. Thus the leader sequences on most vertebrate mRNAs fall in the range of 20 to 100 nucleotides. Note that the mRNAs derived from proto-oncogenes are again atypical, as nearly all of them have very long leader sequences. Also note some extraordinarily long leader sequences (400 to >800 nucleotides) that contain *no* upstream ATG codons; see entries 523, 524 and 650.

The effects of leader length on translational efficiency are just beginning to be explored. There is some evidence that an ATG codon is not recognized efficiently when it occurs close to (within 10 nucleotides of) the cap; that might explain the rare examples of cellular mRNAs in which ribosomes initiate at the first *and second* ATG codons (see entries 143, 297, 330 and footnote *g* in the Appendix). A few viral mRNAs that seem to translate efficiently have leader sequences only 9 or 10 nucleotides long (15-17), but the possibility that some ribosomes reach the second ATG in those mRNAs has not been ruled out. In the case of SV40 late 16S mRNA, the ATG codon that initiates the agnoprotein occurs 10 nucleotides down from the cap and it is clearly recognized inefficiently, despite a favorable context; lengthening the leader sequence by 33 nucleotides seems to improve initiation at the agnoprotein start site, with the result that fewer ribosomes reach the downstream VP1 start site (18). In the case of vaccinia virus, a novel transcriptional maneuver (19) adjusts both the length and the context in a way that favors the efficient translation of late mRNAs.

Whereas a leader sequence that is too short might be deleterious, there is no evidence that long leader sequences are incompatible with efficient translation provided that inhibitory features (notably secondary structure and upstream ATG codons) are avoided. Many of the naturally occurring long 5'-noncoding sequences are G+C rich, however, and therefore secondary structure might negate any advantage of length. From an opposite perspective, the presence of secondary structure in GC-rich leader sequences might *necessitate* that they be long, since length seems to overcome the inhibitory effect of secondary structure in some experimental constructs (M.K., unpublished data).

Errors of note

By comparing two independently derived cDNA sequences for a particular mRNA or by comparing a cDNA with the corresponding genomic sequence, one can spot certain types of errors. The mistakes encountered most frequently when analyzing 5'-noncoding sequences merit comment.

(i) cDNA sequences sometimes correspond, not to the functional mRNA, but to a partially processed precursor that retains an intron in the 5'-noncoding sequence (20-22). Several of the long, ATG-burdened 5'-sequences that have appeared in the literature represent introns that are not present in the mature functional mRNA (23-25). The abundance of upstream ATG codons in the mRNA that encodes a 70K protein associated with Ul RNA (entry 556 in the Appendix) raises the interesting possibility that that cDNA corresponds to an incompletely spliced transcript, and that the splicing machinery--of which Ul snurps are a part--is itself regulated by the efficiency of splicing. With other genes there is indeed experimental evidence for regulation at the level of retaining or removing a 5'-intron (14).

(ii) cDNA sequences sometimes correspond to minor mRNA species that have unusually long (sometimes ATG-afflicted) leader sequences. There are many examples in which Sl nuclease mapping has revealed the bulk mRNA population to have a shorter leader sequence than the longest cDNA (26-32). It is reassuring, therefore, when steps are taken to show that an unusually long leader sequence is really representative of the mRNA population (33, 34).

(iii) Even with Sl mapping, the major transcriptional start site has sometimes been misidentified. For example, the more abundant leader on mouse dihydrofolate reductase mRNA was missed because its shorter length and lower GC content made it less stable (as a DNA·RNA hybrid) than a minor, long leader sequence (35).

(iv) The primer extension technique is error prone, resulting in frequent mistakes in the deduction of 5'-noncoding sequences (36-41; compare 42 with 43 [albumin]; 44 with 45 [ferritin]; 46 with 47 [parathyroid hormone]; 48 with 49 [pyruvate kinase]; 50 with 51 [IGF-II]; and 52 with 53 [X-CGD]. With perverse consistency, such cloning errors generate upstream ATG codons that are not really present in the mRNA.

ACKNOWLEDGEMENT

Research funds were provided by a grant from the National Institutes of Health (GM33915). I thank the staff of Langley Library for their help.

REFERENCES

1. Kozak, M. (1981) Nucl. Acids Res. 9, 5233-5252.
2. Kozak, M. (1984) Nucl. Acids Res. 12, 857-872.
3. Kozak, M. (1986) Cell 44, 283-292.
4. Kozak, M. (1987) J. Mol. Biol. 196, 947-950.
5. Trifonov, E.N. (1987) J. Mol. Biol. 194, 643-652.
6. Heidecker, G. and Messing, J. (1986) Ann. Rev. Plant Physiol. 37, 439-466.
7. Cavener, D.R. (1987) Nucl. Acids Res. 15, 1353-1361.
8. Hamilton, R., Watanabe, C.K. and de Boer, H.A. (1987) Nucl. Acids Res. 15, 3581-3593.
9. Kozak, M. (1986) Proc. Natl. Acad. Sci. USA 83, 2850-2854.
10. Propst, F., Rosenberg, M., Iyer, A., Kaul, K. and Vande Woude, G.F. (1987) Mol. Cell. Biol. 7, 1629-1637.
11. Ratner, L., Thielan, B. and Collins, T. (1987) Nucl.Acids Res. 15, 6017-6036.
12. Kozak, M. (1986) Cell 47, 481-483.
13. Kozak, M. (1987) Mol. Cell. Biol. 7, in press.

14. Gaul, U., Seifert,E., Schuh, R. and Jäckle, H. (1987) Cell 50, 639-647.
15. Dasgupta, R., Shih, D., Saris, C. and Kaesberg, P. (1975) Nature 256, 624-628.
16. Rose, J.K. (1980) Cell 19, 415-421.
17. Collins, P.L. and Wertz, G.W. (1985) J. Virol. 54, 65-71.
18. Grass, D.S. and Manley, J.L. (1987) J. Virol. 61, 2331-2335.
19. Schwer, B., Visca, P., Vos, J. and Stunnenberg, H. (1987) Cell 50, 163-169.
20. McPhaul, M. and Berg, P. (1987) Mol. Cell. Biol. 7, 1841-1847.
21. Larhammar, D., Hammerling, U., Rask. L., and Peterson, P. (1985) J. Biol. Chem. 260, 14111-14119.
22. Ueda, K., Clark, D., Chen, C., Roninson, I., Gottesman, M.M. and Pastan, I. (1987) J. Biol. Chem. 262, 505-508.
23. Wells, D. and Kedes, L. (1985) Proc. Natl. Acad. Sci. USA 82, 2834-2838.
24. Li, S., Tiano, H., Fukasawa, K., Yagi, K., Shimizu, M., Sharief, F., Nakashima, Y. and Pan, Y.E. (1985) Eur. J. Biochem. 149, 215-225.
25. Fukasawa, K.M. and Li, S. (1986) Biochem. J. 235, 435-439.
26. Tsuchiya, M., Kaziro, Y. and Nagáta, S. (1987) Eur. J. Biochem. 165, 7-12.
27. Shahan, K., Gilmartin, M. and Derman, E. (1987) Mol. Cell. Biol. 7, 1938-1946.
28. Rixon, M., Chung, D. and Davie, E. (1985) Biochemistry 24, 2077-2086.
29. Persico, M.G., Viglietto, G., Martini, G., Toniolo, D., Paonessa, G., Moscatelli, C., Dono, R., Vulliamy, T., Luzzatto, L. and D'Urso, M. (1986) NAR 14, 2511-2522.
30. Kobilka, B., Frielle, T., Dohlman, H., Bolanowski, M., Dixon, R., Keller, P., Caron, M. and Lefkowitz, R.J. (1987) J. Biol. Chem. 262, 7321-7327.
31. Akeson, A., Wiginton, D., States, J.C., Perme, C., Dusing, M. and Hutton, J.J. (1987) Proc. Natl. Acad. Sci. USA 84, 5947-5951.
32. Dente, L., Pizza, M.G., Metspalu, A. and Cortese, R. (1987) EMBO J. 6, 2289-2296.
33. Conboy, J., Kan, Y.W., Shohet, S. and Mohandas, N. (1986) Proc. Natl. Acad. Sci. USA 83, 9512-9516.
34. Peralta, E., Winslow, J., Peterson, G., Smith, D., Ashkenazi, A., Ramachandran, J., Schimerlik, M. and Capon, D.J. (1987) Science 236, 600-605.
35. Sazer, S. and Schimke, R.T. (1986) J. Biol. Chem. 261, 4685-4690.
36. Ruppert, S., Scherer, G. and Schütz, G. (1984) Nature 308, 554-557.
37. Auron, P., Webb, A., Rosenwasser, L., Mucci, S., Rich, A., Wolff, S.M. and Dinarello, C.A. (1984) Proc. Natl. Acad. Sci. USA 81, 7907-7911.
38. Ahn, T.G., Cohn, D.V., Gorr, S.U., Ornstein, D.L., Kashdan, M.A. and Levine, M.A. (1987) Proc. Natl. Acad. Sci. USA 84, 5043-5047.
39. Hall, L., Craig, R.K., Edbrooke, M.R. and Campbell, P.N. (1982) Nucl. Acids Res. 10, 3503-3515.
40. Claesson, L., Larhammar, D., Rask, L. and Peterson, P.A. (1983) Proc. Natl. Acad. Sci. USA 80, 7395-7399.
41. Daddona, P.E., Shewach, D.S., Kelley, W.N., Argos, P., Markham, A.F. and Orkin, S.H. (1984) J. Biol. Chem. 259, 12101-12106.
42. Lawn, R.M., Adelman, J., Bock, S.C., Franke, A.E., Houck, C.M., Najarian, R.C., Seeburg, P.H. and Wion, K.L. (1981) Nucl. Acids Res. 9, 6103-6114.
43. Minghetti, P., Ruffner, D., Kuang, W-J., Dennison, O., Hawkins, J., Beattie, W.G. and Dugaiczyk, A. (1986) J. Biol. Chem. 261, 6747-6757.
44. Dörner, M.H., Salfeld, J., Will, H., Leibold, E.A., Vass, J.K. and Munro, H.N. (1985) Proc. Natl. Acad. Sci. USA 82, 3139-3143.
45. Santoro, C., Marone, M., Ferrone, M., Costanzo, F., Colombo, M., Minganti, C., Cortese, R. and Silengo, L. (1986) Nucl. Acids Res. 14, 2863-2876.
46. Kronenberg, H.M., McDevitt, B., Majzoub, J., Nathans, J., Sharp, P.A., Potts, J.T., Jr. and Rich, A. (1979) Proc. Natl. Acac, Sci. USA 76, 4981-4985.
47. Weaver, C.A., Gordon, D.F. and Kemper, B. (1982) Mol. Cell. Endocrinology 28, 411-424.
48. Inoue, H., Noguchi, T. and Tanaka, T. (1986) Eur. J. Biochem. 154, 465-469.
49. Cognet, M., Lone, Y.C., Vaulont, S., Kahn, A. and Marie, J. (1987) J. Mol. Biol. 196, 11-25.

50. Soares, M.B., Ishii, D.N. and Efstratiadis, A. (1985) Nucl. Acids Res. 13, 1119-1134.

51. Soares, M.B., Turken, A., Ishii, D., Mills, L., Episkopou, V., Cotter, S., Zeitlin, S. and Efstratiadis, A. (1986) J. Mol. Biol. 192, 737-752.

52. Royer-Pokora, B., Kunkel, L., Monaco, A., Goff, S., Newburger, P., Baehner, R., Cole, F.S., Curnutte, J.T. and Orkin, S.H. (1986) Nature 322, 32-38.

53. Orkin, S.H. (1987) Trends in Genetics 3, 207.

54. Kozak, M. (1983) Cell 34, 971-978.

APPENDIX

Entry No.	Messenger RNA/source[a]	Leader length[b]	Sequence flanking initiator codon	Upstream ATGs[c] w	w/t	s	s/t	References[d]
	Acetylcholine receptors							
001	α nicotinic, hu muscle		ctccggtagcccATGg					Noda '83 Nature 305,818
002	α nicotinic, rat neural	225	cgggtcttagacATGg					Boulter '86 Nat 319,368
003	β nicotinic, bo muscle		cgcgggcgggctATGa					Tanabe '84 EJB 144,11
004	γ nicotinic, hu muscle		agctgaggcaccATGc					Shibahara '85 EJB 146, 15
005	ε nicotinic, bo muscle		ccagacagcggaATGg					Takai '85 Nat 315, 761
006[e]	atrial muscarinic, po	270, 400	agagaacacaaaATGa		3			Peralta '87 Sci 236,600
007	cerebral muscarinic, po	>444	ocacccagcaccATGa	1	2		1	Kubo '86 Nature 323,411
008	α_1 acid glycoprotein, hu	36	cctggtctcagtATGg					Dente '87 EMBO 6, 2289
009	α_1 acid glycoprotein, rat	40	agtgtcttcggcATGg					Liao '85 MCB 5,3634
	Actins							
010	α-skeletal, mu	70 *	aaactagacaccATGt					Hu '86 MCB 6,15
011	α-skeletal, ch	73 *	acagccagcaacATGt					Fornwald '82 NAR 10,3861
012	α-skeletal, Xp	>60	ccagcctcaaacATGt					Stutz '86 JMB 187,349
013	α-cardiac, Xp	>53 *	taacctgccactATGt					"
014	α-cardiac, ch	60 *	ctatcagccaagATGt					Chang '85 NAR 13,1223
015	α-smooth muscle, hu		gcagctccagctATGt					Ueyama '84 MCB 4,1073
016	α-smooth muscle, ch	88 *	ttgacatagaagATGt					Carroll '86 JBC 261,8965
017	β-cytoplasmic, hu	84 *	cgccagctcaccATGg					Ng '85 MCB 5,2720
018	β-cytoplasmic, rat	80 *	caccagttcgccATGg					Nudel '83 NAR 11,1759
019	β-cytoplasmic, ch	96 *	ccacagccagccATGg					Kost '83 NAR 11,8287
020	γ-cytoplasmic, hu	>73	ctgccggtcgcaATGg					Erba '86 NAR 14,5275
021	3rd cytopl isoform, ch	49 *	gcaggcccaatcATGg					Bergsma '85 MCB 5,1151
022	APRT, hu		tcttcgcacgccATGg					Broderick '87 PNAS 84,3349
023	APRT, mu	60	acgcacgcggccATGt					Dush '85 PNAS 82,2731
024	Adenosine deaminase, hu	95	cacgagggcaccATGg					Ingolia '86 MCB 6,4458
025	Adenosine deaminase, mu	∿135	acgctoggaaccATGg					"
026	AdoHcy hydrolase, rat		gacttcgccagcATGg					Ogawa '87 PNAS 84,719
027	Adenylate kinase, ch	>57	cacagcagcagcATGt					Kishi '86 JBC 261,2942
028	Adipocyte P2, mu	67	aaggtttacaaaATGt					Hunt '86 PNAS 83,3786
029	Adrenodoxin, bo	>164	ccccgacaggctATGg					Okamura '85 PNAS 82,5705
030	Albumin, hu serum	39	gcctttggcacaATGa					Minghetti '86 JBC 261, 6747
031	Albumin, ch serum	41	taatctgcagccATGa					Haché '83 JBC 258,4556
032	ADH, α subunit, hu class I		gacagaatcaacATGa					Ikuta '86 PNAS 83,634
033	ADH, β subunit, hu "	70	gacagaaacgacATGa					Duester '86 JBC 261,2027
034	ADH, γ subunit, hu "	>80	gacagaatcaatATGa	1				Höög '86 EJB 159,215
035	ADH-AA, mu liver	>101	aggacagacggcATGa				1	Edenberg '85 PNAS 82,2262
036	ALDH, hu liver	>30	tagcccgctgcgATGt					Braun '87 NAR 15,3179
037	Aldolase A, rat	66, 116*	gccaccggcaccATGc					Joh '86 JMB 190,401
038	Aldolase B, rat	81 *	gtacctgtcatcATGg					Tsutsumi '85 JMB 181,153
039	Aldolase B, ch	72 *	caataagtcaccATGa					Burgess '85 JBC 260,4604
040	Aldolase C, mu	>60	acaactgtcatcATGc					Paolella '86 EJB 156,229
041	ALP, hu intestinal		tgcccccaagacATGc					Henthorn '87 PNAS 84,1234
042	ALP, hu liver/bone	>176	ttggggtgcaccATGa					Weiss '86 PNAS 83,7182
043	ALA-D, hu	>66	ctggcccacgccATGc	1				Wetmur '86 PNAS 83,7703
044	ALA-D, rat	>48	ccggcccccaccATGc					Bishop '86 NAR 14,10115
045	5-ALV synthase, ch	81	gcaggaggaaggATGg					Maguire '86 NAR 14,1379
046	α-amylase, hu salivary	220 *	cttcaaagcaaaATGa					Nishide '86 Gene 41,299
047	α-amylase, mu salivary	95 *	cagcatagcaaaATGa		1			Hagenbüchle '81 Nat 289,643

No.				w w/t	s s/t	
048	α-amylase, mu pancreatic	17	cttcaaagcaaaATGa			Hagenbüchle'80 Cell 21,179
049	Amyloid-A4 (Alzheimer's)	>146	cgcagggtcgcgATGc			Kang '87 Nature 325,733
050	Amyloid (SAA2), mu serum	34. *	gacaccagcaggATGa			Lowell '86 JBC 261,8442
051	Androgen-BP, 45K, rat	>33	cagctgctaactATGg			Joseph '87 PNAS 84,339
052	And.-induced RP2, mu	>40	aggacgccggccATGa			King '86 NAR 14,5159
053	And-induced S-protein, rat	>51	ttttctggcaagATGa			McDonald '84 EMBO 3,2517
054	Angiotensinogen, rat	61 *	cacagatccgtgATGa			Tanaka '84 JBC 259,8063
055	Antifreeze protein, flounder	49	caagttctcaaaATGg			Davies '84 JBC 259,9241
056	" ocean pout	>57	tcagccacagccATGa			Li '85 JBC 260,12904
057	Arginase, hu liver	>56	aagtgtcagagcATGa			Haraguchi '87 PNAS 84, 412
058	Arginase, rat liver	>26	ccctggatgagcATGa	1		Kawamoto '87 JBC 262, 6280
059	Arginosuccinate lyase hu	>114	gaaccgcccaacATGg			O'Brien '86 PNAS 83,7211
060	Arginosucc. synthase, hu	>75	atcccagacgctATGt			Bock '83 NAR 11, 6505
061	AspAT, mu mitochondrial	>88	ttaccgcccaccATGg			Obaru '86 JBC 261,16976
062	AspAT, mu cytoplasmic	>54	cattctgtcgcgATGg			"
063	AspAT, ch mitochondrial		cacgctgccgccATGg			Jaussi '85 JBC 260,16060
064	ATP/ADP carrier, hu	>70	ccttctttcaacATGa			Battini '87 JBC 262, 4355
065	ATPase-Ca^{2+} ra slow twch	>129	gcccccgcagccATGg			MacLennan'85 Nat 316,696
066	ATPase-Ca^{2+} ra fast twch	>81	gaagggagcgcaATGg			Brandl '86 Cell 44,597
067	ATPase(Na^{+}K^{+})-α, rat brain	>237	agcgccgccaccATGg			Shull '86 Biochem 25, 8125
068	" " αIII "	>140	ggagccgccaagATGg			"
069	" " β, human	>120	cccgccatcgccATGg	2		Kawakami '86 NAR 14,2833
070	" " β rat kidny	>460	tgagcagacaccATGg			Young '87 JBC 262,4905
071	" " α sh kidney	>264	accaccgccgctATGg			Shull '85 Nature 316, 691
072	" " β sh kidney	>528	tgacccgccaccATGg			Shull '86 Nature 321, 429
073	" (H^{+}K^{+}), rat stomach	>206	cacctagccaccATGg	5	4	Shull '86 JBC 261, 16788
074	ATPase, Xp mitochondrial		caagccgcagtcATGt			Weeks '87 PNAS 84,2798
075	Atrial natriuretic factor	90	acccacgccagcATGg			Argentin '85 JBC 260, 4568
076	Avidin, ch	>43	cctgctgcagagATGg			Gope '87 NAR 15, 3595
077	BPGM, hu erythrocyte	>110	tcagccatcagtATGt	1		Joulin '86 EMBO 5, 2275
078	Bone Gla protein, rat	>48	ctagcagacaccATGa			Celeste '86 EMBO 5, 1885
079	Brain S100-α protein, bo	>89	gtaagcttcgagATGg			Kuwano '86 FEBS 202, 97
080	Brain S100-β protein, rat	>120	ggagcctccgggATGt			Kuwano '84 NAR 12, 7455
081	Brain spcif. gene O-44, rat	66 +	taggccgccgagATGg	±		Tsou '86 MCB 6, 768
082	C-reactive protein, hu	104	caggacgtgaccATGg			Lei '85 JBC 260, 13377
083	Caerulein, Xp	>50	ccttctgaaagcATGt			Richter '86 JBC 261, 3676
084	Calcitonin, hu	>74	cagagaggtgtcATGg			Jonas '85 PNAS 82, 1994
085	Calcitonin, rat	132	cagggaggcatcATGg			Amara '84 MCB 4, 2151
086	Ca^{2+}binding protein, ch	>102	tgcaccccaacATGa			Hunziker '86 PNAS 83, 7578
087	Ca^{2+} " rat brain	>130	agccgctgcaccATGg			Yamakuni '86 NAR 14, 6768
	Calmod. family Ca^{2+}binding proteins:					
088	Calmodulin, rat	85	ttcgctcgcaccATGg			Nojima '87 JMB 193, 439
089	Calmodulin pRCM3, rat	>70	agcccttgcagcATGg			Nojima '87 MCB 7, 1873
090	Calmodulin, ch	91	ggccgagccaccATGg			Putkey '83 JBC 258, 11864
091	Calmodulin, Xp	>70	aactattccgaaATGg			Chien '84 MCB 4, 507
092	Oncomodulin, rat	97	gcgggacagaaaATGa			Gillen '87 JBC 262, 5308
093	Parvalbumin, rat	73 *	ccaagttgcaggATGt			Epstein '86 JBC 261, 5886
094	Calpain, po	>90	tgagtcacagccATGt	1		Sakihama '85 PNAS 82, 6075
095	Calcyclin (S100-related) hu	65 *	cagccctcagccATGg			Ferrari '87 JBC 262, 8325

No.					w	w/t	s	s/t	
096	Carbamyl-P-synthetase rat	140		aacatcttcaaaATGa		1			Lagacé '87 JBC 262,10415
097	Carbonic anhydrase I, hu			cagtagaagataATGg					Barlow '87 NAR 15, 2386
098	Carbonic anhydrase II, mu	60		accggcgtgaccATGt					Venta '85 JBC 260, 12130
099	Carbonic anhydrase III, hu	>43		aggaaggcgaccATGg					Lloyd '86 Gene 41, 233
100	Carbonic anhydrase II, ch	39		ggccggcgcaccATGt					Yoshihara '87 NAR 15, 753
101	Cartilage-link protein ch	>135		gtgactgtgaagATGa					Deak '86 PNAS 83, 3766
102	α-Casein, rat	62	*	atcttagcaaccATGa					Yu-Lee, 86 NAR 14, 1883
103	β-Casein, rat	52	*	gacttgacagccATGa					Jones '85 JBC 260, 7042
104	β-Casein, mu	>55		gacttgacagccATGa					Yoshimura '86 NAR 14, 8224
105	γ-Casein, rat	56	*	gatcaagtaaccATGa					Hobbs '82 NAR 10, 8079
106	Catalase, hu kidney	70		aaaccgcacgctATGg					Quan '86 NAR 14, 5321
107	Catalase, rat liver	>83		caatcctacaccATGg					Furuta '86 PNAS 83, 313
108	Cathepsin B, hu	>195		ctggcttccaacATGt		1			Chan '86 PNAS 83, 7721
109	Cathepsin D, hu	>51		gccgccgccgccATGc					Faust '85 PNAS 82, 4910
110	Cholecystokinin, hu	64	*	aatccaaaagccATGa					Takahashi '86 Gene 50, 353
111	Cholecystokinin, rat	59	*	gcatccgaagatATGa					Deschenes '85 JBC 260, 1280
112	Chromogranin, bo	180		ccoggcttcgccATGc					Benedum '86 EMBO 5, 1495
113	proChymosin, bo	25		cccagatccaagATGa					Hidaka '86 Gene 43, 197
114	Chymotrypsingen-B, rat	22		ttgaccagcaccATGg					Bell '84 JBC 259, 14265
	Coagulation factors & modulators:			see also Fibrinogen, Inhibitors, PA, Thrombospondin					
115	Factor VII, hu	>35		agagatttcatcATGg					Hagen '86 PNAS 83, 2412
116	Factor VIII, hu	170		tagcaataagtcATGc				1	Gitschier '84 Nat 312,326
117	Factor X, bo	>75		aagggccccaccATGg					Fung '84 NAR 12, 4481
118	Factor XIII-α, hu placenta	>84		gtaaagtcaaaaATGt					Grundmann'86 PNAS 83,8024
119	von Willebrand factor, hu	246		ttgcaggggaagATGa		1		1	Verwèij '86 EMBO 5 ,1839
120	Protein C, hu	75	*	agtgcctccagaATGt			1		Plutzky '86 PNAS 83, 546
121	Protein S, hu	>108		cgcgccttcgaaATGa					Hoskins '87 PNAS 84, 349
122	Protein S, bo	>35		gcccgtttcgccATGa					Dahlbäck '86 PNAS 83, 4199
123	proCollagen, α2(I), ch	133		tagcaagtagacATGc		2			Vogeli '81 PNAS 78, 5334
124	proCollagen, α1(I), ch	>100		taatatttagacATGt		2			Yamada '83 JBC 258, 14914
125	proCollagen, α1(II), rat	155		tcgcggtgagccATGa		1			Kohno '85 JBC 260, 4441
126	Collagenase, hu skin	>68		acaaaggccagtATGc					Goldberg '86 JBC 261,6600
	Complement components & modulators:								
127	preClr, hu	>51		gggccttgagaaATGt					Journet '86 Bio. J. 240,783
128	preC2, hu	>36		agggaggacaccATGg					Bentley '86 Bio. J. 239,339
129	preC3, hu	>60		tgtcccagcaccATGg					deBruijn '85 PNAS 82,708
130	preC3, mu	56		ttttccttcactATGg					Wiebauer '82 PNAS 79,7077
131	preC4, mu	61		gatcctccagccATGc					Nonaka '86 PNAS 83,7883
132	Factor B, hu	129		ccttccaacgccATGg		1			Wu '87 Cell 48,331
133	Factor 1, hu			aacacctccaacATGa					Catterall'87 BioJ 242,849
134	Decay-accelerate factor hu	>66		acccggcgcgccATGa					Caras '87 Nature 325, 545
135	Cl inhibitor, hu			gtcgccgcccagATGg					Bock '86 Biochem 25, 4292
136	Conalbumin, ch	76		ccctgccccaacATGa					Jeltsch '82 EJB 122,291
137	Corticotropin-RF, sh	>127	*	gcgccccctaacATGc					Furutani '83 Nature 301,537
138	Creatine kinase, rat brain	>29		gccgccgccgccAUGc					Benfield '85 Gene 39, 263
139	Creatine kinase, ra muscle	>50		gacgccgccaccATGc					Putney '84 JBC 259, 14317
140	Creatine kinase-B, ch	>42		gtagggacagccATGc					Hossle '86 NAR 14, 1449
141	αA-Crystallin, ha	68		gccaagaagaacATGg					Heuvel '85 JMB 185, 273
142	αB-Crystallin, ha	43		cacctagccaccATGg					Quax-Jeuken'85 PNAS 82,5819
1439	βA3/A1-Crystallin, mu	56		ccaaccaccaagATGg					Peterson '86 Gene 45,139
144	βB1-Crystallin, ch	>70		ctgaccaccgcgATGt					Hejtmancik'86 JBC 261,982

No.					w	w/t	s	s/t	
145	8B1-Crystallin, rat	38	*	gcatcagaaaccATGt					Dunnen '86 PNAS 83,2855
146	γ-Crystallin, rat	42		caaacaacagccATGg					Moormann '83 JMB 171, 353
147	δ1-Crystallin, ch	86	*	acgtcgtccgaaATGg					Hayashi '85 EMBO 4, 2201
148	Cell cycle ts11 gene, hu	>78		tgtctgagtgctATGa		1			Greco '87 PNAS 84, 1565
149	Cell cycle "cdc2", hu	>140		cattgactaactATGg					Lee '87 Nature 327, 31
150	Cyclin, hu	>118		cactccgccaccATGt					Almendral '87 PNAS 84,1575
151	Cyclophilin, hu T-cell			gtactattagccATGg					Haendler '87 EMBO 6, 947
152	Cystic fibrosis Ag, hu	51		tccgtgggcatcATGt					Dorin '87 Nature 326, 614
153	Cytochrome c, mu		*	ttagaattaaaaATGg					Limbach '85 NAR 13, 617
154	Cytochrome c, ch			ctagtactgacaATGg					Limbach '83 NAR 11, 8931
155	Cytochr. c oxidase-IV, bo	32		ggtggcatcagaATGt					Lomax '84 PNAS 81, 6295
156	NADPH-cyto P450 reductase	70		tgtatcaccaacATGg					Murakami '86 DNA 5, 1
	Cytochrome P-450 proteins:								
157	P$_1$-450, human	122	*	cctacactgatcATGc					Jaiswal '85 NAR 13,4503
158	P-450$_{NP}$, human	>68		aggaaagtagtgATGg					Beaune '86 PNAS 83, 8064
159	P-450c17, human	>41		cacccagccaccATGt					Chung '87 PNAS 84, 407
160	P-450c21, human	9, 53, 118		gggcgtctcgccATGc	±				Higashi '86 PNAS 83, 2841
161	P-450scc, human	75		tgtggggacagcATGc					Chung '86 PNAS 83, 8962
162	P-450 1, rabbit	>43		aggaagaagagaATGg					Johnson '87 JBC 262, 5918
163	P-450(M-1), male rat liver			gagaaggctgccATGg					Yoshioka '87 JBC 262, 1706
164	P-450MC, rat	>72		gccacctagatcATGc					Yabusaki '84 NAR 12, 2929
165	P-450dbl, rat	>51		agcaaggcagccATGg					Gonzalez '87 DNA 6, 149
166	P-450-PCN, rat	>82		agacctgcagggATGg	1				" '85 JBC 260,7435
167	P-450e, rat	30		tacaccaggaccATGg					Mizukami '83 PNAS 80, 3958
168	P-450a, rat			tcactggccactATGc					Nagata '87 JBC 262, 2787
169	P-450s, ch	>39		cctctgcccaccATGg					Hobbs '86 JBC 261, 9444
170	Cytokeratin type I, bo	25		aacagcatcaccATGt					Rieger '85 EMBO 4, 2261
171	Cytokeratin 19, bo	59		tcctgcttcgccATGa					Bader '86 EMBO 5, 1865
172	Cytokeratin Endo A, mu	79		cagacttcaccaATGt					Vasseur '85 PNAS 82, 1155
173	Cytokeratin Endo B, mu	>54		tctccagacaagATGa					Singer '86 JBC 261, 538
174	Cytokeratin-8, Xp	>36		cacagctccaccATGt					Franz '86 PNAS 83, 6475
175	Desmin, ha	81		cacgccgccaccATGa					Quax '85 Cell 43,327
176	Diazepam inhibitor, rat	>117		cacctcgccagtATGt					Mocchetti '86 PNAS 83, 7221
177	DHFR, hu	71		cccgctgctgtcATGg					Chen '84 JBC 259, 3933
178	DHFR, mu	55, 115		cccgctgccatcATGg					Nunberg '80 Cell 19, 355
179	Dihydropteridine reductase	>80		ggcaggagcaggATGg				1	Lockyer '87 PNAS 84, 3329
180	Elastase I, rat pancreatic	22		tctctccacaacATGc					MacDonald '82 Biochem
181	Elastase II, " "	22		cacggacacaccATGA					21, 1453
182	Elastin, hu	>21		tttctccccgagATGg					Indik '87 PNAS 84, 5680
183	Endopeptidase, neutral, ra	>		agattttaggtgATGg					Devault '87 EMBO 6, 1317
184	Endothelial cell GF, hu	>38		agctgctgagccATGg					Jaye '86 Sci 233, 541
185	preproEnkephalin A, hu	130	*	agcgtcaactccATGg					Noda '82 Nature 297, 431
186	preproEnkephalin B, hu			tgctgagacaggATGg					Horikawa '83 Nat 306, 611
187	preproEnkephalin, rat	156	*	accggcagccccATGg					Rosen '84 JBC 259, 14309
188	Enolase, non-neuron, rat	>110		cagaacttcaccATGt					Sakimura '85 NAR 13, 4365
189	Enolase, neuronal, rat	>68		atcccagccatcATGt					Sakimura '85 PNAS 82, 7453
190	Epidermal GF(pre), hu	>436		ctcatcaagattATGc		1		1	Bell '86 NAR 14, 8427
191	Epidermal GF(pre), mu	350		gctaaataaaagATGc					Scott '83 Sci 221, 236
192	Epididymal proteins D, E	>82		gaaaatagaaccATGg					Brooks '86 EJB 161, 13
193	Epoxide hydrolase, rat	175	*	cagtcaggagtcATGt				1	Falany '87 JBC 262, 5924
194	Erythroid membr pr 4.1, hu	>798		aaacacaggaacATGc		3		1	Conboy '86 PNAS 83, 9512
195	Erythroid potentiating, hu	>72		agagaacccaccATGg					Carmichael '86 PNAS 83,2407

No.					w	w/t	s	s/t	
196	Erythropoietin, hu	>180		ccaggcgcggagATGg		1			Jacobs '85 Nature 313, 806
197	Fatty acid binding protein rat liver	48		ctcattgccaccATGa					Sweetser '86 JBC 261, 5553
198	" rat intestine			acagctgacatcATGg					Alpers '84 PNAS 81, 313
199	FA thioesterase, rat	>310		tcaactcacagaATGg					Safford '87 Biochm 26,1358
200	FA thioesterase, duck	>160		aattattcaggaATGg					Poulose '85 JBC 260,15953
201	Ferritin, L-chain, hu	181		tgccaaccaaccATGa					Santoro '86 NAR 14, 2863
202	Ferritin, L-chain, rat	200		tccggatcagccATGa					Leibold '87 JBC 262,7335
203	Ferritin, H-chain, hu	216		ttagtcgccgccATGa					Hentze '86 PNAS 83, 7226
204	Ferritin, H-chain, ch	151		cccgccagcgccATGg					Stevens '87 MCB 7, 1751
205	Ferritin, bullfrog	145		caaaccgctgaaATGg					Didsbury '86 JBC 261, 949
206	α-Fetoprotein, hu	44		taactagcaaccATGa					Gibbs '87 Biochem 26, 1332
207	preFibrinogen, Aα, hu	>57		cccttagaaaagATGt					Kant '83 PNAS 80, 3953
208	preFibrinogen-γ, hu	51		cactcagacatcATGa					Rixon '85 Biochem 24, 2077
209	preFibrinogen-α, rat	58		caatcagaaactATGc	1				Crabtree '85 JMB 185, 1
210	" -β, rat			aaatctgaaaccATGa					Fowlkes '84 PNAS 81, 2313
211	" -γ, rat	44		cacccagacactATGa					Morgan '87 NAR 15, 2774
212	Fibroblast growth factor	>320		cccgcagggaccATGg					Abraham '86 EMBO 5, 2523
213	Fibronectin, hu	267		caccgtctcaacATGc					Dean '87 PNAS 84, 1876
214	FSH, β-chain, bo	>48		caagtgcccaggATGa					Esch '86 PNAS 83, 6618
	Guanine nucleotide binding proteins:								
215	$G_{o\alpha}$, bovine retina	>196		gaaggggccaccATGg					Van Heurs '87 PNAS 84, 3107
216	$G_{i\alpha}$, rat brain	>41		gcggacggcaggATGg					Itoh '86 PNAS 83, 3776
217	$G_{s\alpha}$, rat brain	>246		cccgccgccgccATGg					"
218	$T_{\alpha 1}$, bovine retina	>93		cctgccaggaccATGg					Medynski '85 PNAS 82,4311
219	$\beta_1$subunit, bovine retina	>94		taagattagaagATGa					Fong '86 PNAS 83, 2162
220	γ-subunit, " "	>67		gcttagcagaagATGc					Yatsunami '85 PNAS 82,1936
221	preproGalanin, po	>225		ccgtcgctcaagATGc		1			Rokaeus '86 PNAS 83, 6287
222	Gap junction protein, rat	>31		gaatgaggcaggATGa	1				Paul '86 J Cell Biol 103,123
223	GAP-43, rat neuronal	>60		gaagataccaccATGc					Karns '87 Sci 236, 597
224	preproGastrin, hu	65	*	tctgcagacgagATGc					Ito '84 PNAS 81, 4662
225	Gastrin-releasing, hu	>55		cccgtcgggaccATGc					Spindel '84 PNAS 81, 5699
226	Gelsolin, hu			cgtgtcgccaccATGg					Yin '86 Nature 323, 455
	α-Globin family:								
227	human adult	37		agagaacccaccATGg					Baralle '77 Cell 12,1085
228	human embryonic (ζ)	55		caccctgccgccATGt					Proudfoot '82 Cell 31, 553
229	baboon θ1	193		tccagcgcgggaATGg					Shaw '87 Nature 326, 717
230	mouse adult	32		caggaagaaaccATGg					Baralle '78 Nature 274,84
231	rabbit adult	36		gaaggaaccaccATGg					Baralle '77 Nature 267,279
232	goat embryonic (ζ)	46		tcagctgccaccATGt					Wernke '86 JMB 192, 457
233	duck adult, major (α^A)	36		ggagctgcaaccATGg					Erbil '82 Gene 20, 211
234	chicken embryonic	55		ctctcctgcacaATGg					Engel '83 PNAS 80, 1392
	β-Globin family:								
235	human fetal (γ)	53		agtccagacgccATGg					Slightom '80 Cell 21,627
236	human embryonic (ε)	53		aggcctggcatcATGg					Baralle '80 Cell 21, 621
237	rabbit adult	53		aaaacagacagaATGg					Baralle '77 Cell 10, 549
238	rabbit embryonic (β3)	62		agaccagacatcATGg					Hardison '81 JBC 256,11780
239	chicken adult	77		ccaaccgccgccATGg					Dolan '83 JBC 258, 3983
240	chicken embryonic (ρ)	45		cccgctgccaccATGg					Roninson '81 PNAS 78, 4782
241	Xenopus adult, major	46		tcaactttggccATGg					Patient '83 JBC 258,8521
242	Xenopus larval			tctacagccaccATGg					Banville '83 JBC 258, 7924
243	$\alpha_{2\mu}$Globulin, rat salivary	70		ttccctaccaacATGa					Laperche '83 Cell 32, 453
244	preproGlucagon, hu pancr.	105	*	aagacagcagaaATGa					White '86 NAR 14, 4719
245	preproGlucagon, rat	101	*	cagaataaaaaaATGa					Heinrich '84 JBC 259,14082
246	preproGlucagon, anglerfish	>58		acggttgtaaacATGa					Lund '82 PNAS 79, 345

No.					v	w/t	s	s/t	
247	Glucose-6-P-dehydrog., hu	70	*	atattcatcatcATGg			1		Martini '86 EMBO 5,1849
248	Glucose-regl'd 78K prot,rat	206		agcgccggcaagATGa					Chang '87 PNAS 84, 680
249	Glucose transporter, hu	>180		cgcagcgctgccATGg					Mueckler '85 Sci 229,941
250	Glucose transporter, rat	>207		cgcagcgcggccATGg		7			Birnbaum '86 PNAS 83, 5784
251	β-Glucuronidase, hu	>26		ggaccgggaagcATGg					Oshima '87 PNAS 84, 685
252	γ-Glutamyl transpeptidase	>228		actggacgagccATGa					Laperche '86 PNAS 83, 937
253	Glutathione peroxidase, mu	37		aacatctccagtATGt					Chambers '86 EMBO 5, 1221
	<u>Glutathione-S-transferases</u>:								
254	-subunit 2, hu	>55		gagactgctatcATGg					Board '87 PNAS 84,2377
255	-Ya subunit, rat	64	*	acagttgctgctATGt				1	Pickett '86 PNAS 83,9393
256	-Yb$_1$ subunit, rat	>37		agcgccagaaccATGc					Ding '85 JBC 260, 13268
257	-Yc subunit, rat	>42		gcaattgctgccATGc					Pickett '85 JBC 260, 5820
258	GST-placental, rat	70		tacgcagcagctATGc					Okuda '87 JBC 262, 3858
259	GAPDH, hu	>75		cgctcagacaccATGg					Tso '85 NAR 13,2485
260	GAPDH, rat	71		ctcatagacaagATGg					Fort '85 NAR 13,1431
261	GAPDH, rat	57	*	tataaaggcgagATGg					Stone '85 PNAS 82,1628
262	Glycero-P-dehydrogenase, mu	21		gcaagcagcaccATGg					Ireland '86 JBC 261,11779
263	Glycogen phosphorylase, hu	>113		gccgccccagccATGg					Newgard '86 PNAS 83, 8132
264	Gonadotropin-β, salmon			tgagtccttcagATGt					Trinh '86 EJB 159,619
265	Growth hormone, hu	60		cacctagctgcaATGg					DeNoto '81 NAR 9, 3719
266	Growth hormone, rat	60		cactgagtggcgATGg					Page '81 NAR 9,2087
267	Growth hormone, salmon	>64		ttaagagtaaaaATGg					Sekine '85 PNAS 82, 4306
268	GH-releasing factor, hu	91	*	cccgggtgaaggATGc					Mayo '85 PNAS 82,63
269	Haptoglobin, hu	30		agaccaaccaagATGa					Bensi '85 EMBO 4, 119
270	Heat shock 70K, hu	212		gcagggaaccgcATGg					Hunt '85 PNAS 82,6455
271	Heat shock 70K, hu	119		gaagcttcagccATGc					Voellmy '85 PNAS 82,4949
272	Heat shock 27K, hu	>40		gagtcagccagcATGa					Hickey '86 NAR 14,4127
273	Heat shock 73K, rat	80	*	acgcaagcaaccATGt					Sorger '87 EMBO 6,993
274	Heat shock 70K, ch	111		gaatctatcatcATGt					Morimoto '86 JBC 261,12692
275	Heat shock 108K, ch	101		ggccgcggcatcATGa					Kulomaa'86 Biochem 25,6244
276	Heat shock 70K, trout	>60		ttattcggtaacATGt					Kothary '84 MCB 4,1785
277	Heat shock 70K, Xp	124		aggagcgcaaatATGg					Bienz '84 EMBO 3,2477
278	Helix-destabilizing, rat	>28		catcctaccgtcATGt					Cobianchi'86 JBC 261,3536
279	Heme oxygenase, rat	128		cccagtcccgcgATGg					Muller '87 JBC 262,6795
280	β-Hexosaminidase, α-chain	>168		gaccagcgggccATGa					Myerowitz '85 PNAS 82,7830
281	HMG-14, hu	>145		ccccgccgccagATGc					Landsman '86 JBC 261,16082
282	HMG-17, hu	>88		gccgccgccaccATGc					Landsman '86 JBC 261,7479
283	His-rich glycoprot., hu	>120		tggtttaacaaaATGa		3		2	Koide '86 Biochem 25,2220
	<u>Histocompatibility antigens</u> (MHC):								
284	Class I(hu): HLA-Bw58			tcagacgccgagATGc					Ways '85 JBC 260,11924
285	Class II(hu): DR(α-chain)	64		cccaagaagaaaATGg					Schamboeck '83 NAR 11,8663
286	Class II(hu): DR(β-chain)	>30		ctgttctccagcATGg					Tieber '86 JBC 261,2738
287	Class II(hu): DC-3β	58		ttcgtctcaattATGt					Boss '84 PNAS 81, 5199
288	Class II(hu): SB(DP)-α			agaccccacaacATGc					Lawrance '85 NAR 13,7515
289	Class I(mu): H-2K^b			tcagtcgtcagcATGg					Kimura '86 Cell 44,261
290	Class I(mu): H-2K^d	>27		ccgaccagtgcgATGg					Lalanne '83 NAR 11,1567
291	Class I(mu): *Tla* gene T3^b			gatttccctaacATGa					Pontarotti'86 PNAS 83,1782
292	Class II(mu): A(β-chain)	>36		tgtgccttagagATGg					McDevitt '87 PNAS 84,2435
293	Class II(mu): A(β2-chain)			gtcccctccagaATGg					Peterson '85 JBC 260,14111
294	Class II(mu): E(α-chain)	48		cccaagaagaaaATGg					Mathis '83 Cell 32,745
295	Class II(mu): E(β-chain)	52		ctctcctgcagcATGg					Saito '83 PNAS 80,5520
296	Class II(mu): E(β2-chain)			ctctcctcaagcATGg					Braunstein '86 EMBO 5,2469
2979	Class II-assoc'd Ia(In), hu	58		cagaagccagtcATGg					Strubin '86 Cell 47, 619

No.				w w/t	s s/t	
298	Histone H1, hu		tttcttgccaccATGt			Carozzi '84 Sci224,1115
299	" H2a, hu	46	tcagaagtagttATGt			Zhong '83 NAR 11, 7409
300	" H2b, hu	40	ctagacagtgctATGc			"
301	" H3, hu	37	tgtggttttgccATGg			"
302	" H3, hu	27	gcagttctgcgaATGg			Clark '81 NAR 9,1583
303	" H3.3, hu	111 *	ggtctctgtaccATGg			Wells '87 NAR 15,2871
304	" H4, hu	36	ttgctcgtcgtcATGt			Heintz '81 Cell 24,661
305	" H2b, mu	41	tctctgttcactATGc			Sittman '83 NAR 11, 6679
306	" H3.1, mu	28	cggttacttgccATGg			"
307	" H3.2, mu	21	ttctttgtagaaATGg			"
308	" H1, ch embry.	38	acgtccgtcaccATGt			Sugarman '83 JBC 258, 9005
309	" H2A.1, ch	146	tcagtcgctgcgATGt			D'Andrea '81 NAR 9,3119
310	" H2A.F, ch embry	70	ggcggcggcaccATGg			Harvey '83 PNAS 80, 2819
311	" H2B, ch	36	ggagagttcgacATGc			Grandy '82 JBC 257, 8577
312	" H3.3A, ch	*	gtcggcagcagcATGg			Brush '85 MCB 5,1307
313	" H3.3B, ch	>105 *	aagtgagaaaaaATGg			Dodgson '87 NAR 15,6294
314	" H4, ch embry.	26	caggctctcggcATGt			Sugarman '83 JBC 258, 9005
315	" H5, ch erythrocyte	109	gaagcggcggccATGa			Krieg '83 NAR 11, 619
316	" H1, Xp	28	tttacttcaaagATGa			Turner '83 NAR 11, 4093
317	" H2A, Xp	47	agoacagtaatcATGt			Moorman '82 FEBS 144,235
318	" H2B, Xp	35	agcagcacaattATGc			"
319	" H3, Xp		aactgatacactATGg			Moorman '81 FEBS 136, 45
320	" H4, Xp	28	gctcaagaaagaATGt			"
321	Hydroxyindole O-MeTr'ase,bo	>120	cccagaaggaagATGt	1		Ishida '87 JBC 262, 2895
322[h]	HMG-CoA reductase, hu	73 to 105*	tctgtagctacaATGt			Luskey '87 MCB 7, 1881; and " '85 JBC 260,10271
323	HMG-CoA synthase, hu	63 or 122	tgctctttcaccATGc			Gil '87 PNAS 84, 1863
324	HPRT, hu	100 to 170	gccggctccgttATGg			Patel '86 MCB 6, 393
325	HPRT, mu	90, 118	accggtcccgtcATGc			Melton '84 PNAS 81, 2147
326	Ig L-chain, kappa-II, hu		caccttctcacaATGa			Klobeck '85 NAR 13, 6499
327	Ig L-chain, kappa-III, hu		cccagaggaaccATGg			"
328	Ig L-chain, kappa-IV, hu	25	aggggcagcaagATGg			Marsh '85 NAR 13, 6531
329	Ig H-chain, IgE, hu	>56	ccgttcctcaccATGg			Kenten '82 PNAS 79, 6661
330[g]	Ig L-chain, kappa, mu	18	catcacaccagcATGg			Kelley '82 Cell 29, 681
331	Ig L-chain, lambda$_I$, mu	40	ggtttgtgaattATGg			Picard '83 PNAS 80, 417
332	Ig H-chain, mu	>45	agtcctgtcactATGa			Early '80 Cell 19,981
333	Ig L-chain, ch		tgggattccgccATGg			Reynaud '87 Cell 48,379
334	Inhibin, A-subunit, hu	>144	gccaggtgagctATGg	1	1	Mayo '86 PNAS 83, 5849
335	Inhibin, A-subunit, bo	>60	gccaggggagctATGt			Forage '86 PNAS 83,3091
	Inhibitors: see also EPA, Kininogen, Lipocortin, Macro- and Microglobulins					
336	anti-Protein C, hu	>46	agaacatccaccATGc			Suzuki '87 JBC 262, 611
337	α1-antiChymotrypsin, hu		gcagagttgagaATGg			Chandra '83 Biochem22,5055
338	anti-Elastase, hu	17	cctgccttcaccATGa			Stetler '86 NAR 14, 7883
339	anti-placental PA, hu	>55	cagattgaaacaATGg			Ye '87 JBC 262, 3718
340	antiThrombin III, hu	~70	agattagcggccATGt			Bock '82 NAR 10, 8113
341	{α$_1$-antiTrypsin, hu liver macrophage	49 * 520 *	agtgaatcgacaATGc	(1)	(1)	Ciliberto '85 Cell 41,531 Perlino '87 EMBO 6, 2767
342	Insulin, hu	59 *	tgtccttctgccATGg			Bell '80 Nature 284, 26
343	Insulin-I, rat	57 *	cattgttccaacATGg			Lomedico '79 Cell 18,545
344	Insulin, gp	60 *	catcctttcatcATGg			Chan '84 PNAS 81, 5046
345[i]	Insulin, ch	*	ccccagctcatcATGg	?		Perler '80 Cell 20, 555
346	Insulin, anglerfish	>85 *	ttctactgcagcATGg			Hobart '80 Sci 210,1360
347	Insulin, salmon	>71	cctaccatcaccATGg			Sorokin '82 Gene 20, 367
348	Insulin-like GF I, hu	>180	acttcagaagcaATGg	1	1	Rotwein '86 PNAS 83, 77
349[j]	Insulin-like GF II, rat	100 *	cttcaggtaccaATGg			Soares '86 JMB 192, 737
350	Integrin, ch	96	cggcgcgcggccATGg			Tamkun '86 Cell 46, 271
351[b]	Interferon-α2, hu(LeIF A)	~70	gcaacatctacaATGg			Lawn '81 PNAS 78, 5435
352	Interferon-α, mu	~70	gcaacactcaccATGg			Shaw '83 NAR 11, 555

No.					w	w/t	s	s/t	
353	Interferon-α, rat	∿72		gccacatttgccATGg					Dijkema '84 NAR 12, 1227
354	Interferon-α, bo	∿70		tcaaggtccccgATGg					Capon '85 MCB 5,768
355	Interferon-β, hu fibroblast	75		cgtgttgtcaacATGa					Ohno '81 PNAS 78,5305
356	Interferon-$β_2$, hu	63		aggagcccagctATGa					Haegeman'86 EJB 159,625
357	Interferon-γ, hu immune	130		ctctcggaaacgATGa					Gray '82 Nature 298, 859
358	Interferon-γ, rat	110		agctctgagacaATGa					Dijkema '85 EMBO 4, 761
359	IFN-induced gene 6-16, hu	106		gcgcgcgccaccATGc					Kelly '86 EMBO 5, 1601
360	IFN-induced ISG-54K, hu	75		tgcagtgcaaccATGa					Levy '86 PNAS 83, 8929
361	IFN-induced 15K prot. hu	>75		cagcccacagccATGg					Blomstrom '86 JBC 261,8811
362	Involucrin, hu	62	*	gtagcttctaagATGt					Eckert '86 Cell 46, 583
364	Keratin-I, 50K, hu	60		ctcctctgcaccATGa					Marchuk '85 PNAS 82,1609
365	Keratin, 67K, hu	47		ctctaagtcaacATGa					Johnson '85 PNAS 82, 1896
366	Keratin-II, 56K, hu			atctctggaaccATGg					Tyner '85 PNAS 82, 4683
367	Keratin-I, 47K, mu	>116		cctctctcagccATGa		1			Knapp '86 NAR 14, 751
368	Keratin-I, 59K, mu	25		cactacaccaccATGt					Krieg '85 JBC 260, 5867
369	Keratin B2A, sh	50		actcctgacaccATGg					Powell '83 NAR 11, 5327
	(see also #170-174)								
370	protein Kinase C, α/β, ra	>221		cccgcgcgcaagATGg					Ohno '87 Nature 325, 161
371	protein Kinase C, γ, ra	>204		ttggggggaccATGg					"
372	protein Kinase, cAMP-alt	>50		atcgccccagtcATGg					Showers '86 JBC 261,16288
373	" " II(Ca^{2+}), rat	∿200		atcgccaccgccATGg					Bennett '87 PNAS 84, 1794
374	preKininogen, hu	130, 154,184		gattgttagatcATGa					Kitamura'85 JBC 260,8610
375	α-Lactalbumin, hu	∿26		ggggtagccaaaATGa					Hall '87 Bioch.J.242,735
376	LDH-A, hu	>99	*	tccaagtccaatATGg					Tsujibo '85 EJB 147, 9
377	LDH-C, mu	>54		gtaaggctcaacATGt					Sakai '87 Bioch.J.242,619
378	Lamin C, hu	>200		aacctgccggccATGg					Fisher '86 PNAS 83, 6450
379	L-C acyltransferase, hu	24		accagggctggaATGg					McLean '86 NAR 14, 9397
380	Lens MIP, bo	>50		atccccctgccATGt					Gorin '84 Cell 39, 49
381	Leukocyte adhesion pr. β, hu	>72		acaccgagggacATGc					Kishimoto '87 Cell 48,681
382	Lipase, rat hepatic	>15		aagacgagagacATGg					Komaromy'87 PNAS 84,1526
383	Lipase, rat lingual	48		tagcagtacaagATGt					Docherty '85 NAR 13, 1891
384	Liprotein lipase, hu	>174		acgcgccccgagATGg					Wion '87 Sci 235, 1638
385	Lipid binding protein, mu	65		aaggtttacaaaATGt					Phillips '86 JBC 261,10821
386	Lipocortin-II, hu	>50		gcttccttcaaaATGt					Huang '86 Cell 46,191
387	apoLipoprotein A-I, rat	40	*	acatccttcaggATGa					Haddad '86 JBC 261,13268
388	apoLipoprotein A-I, ch			ttcagcgcgaagATGa					Lusis '87 JBC 262, 7058
389	apoLipoprotein A-II, hu	58	*	actgttaccaacATGa					Shelley '85 JMB 186, 43
390	apoLipoprotein A-II, mu	>41		tagtctgccatcATGa					Kunisada '86 NAR 14,5729
391	apoLipoprotein II,VLD, ch	77	*	taccaacaaaccATGg					AB '83 NAR 11, 2529
392	apoLipoprotein A-IV, mu	91		tgaggagccaggATGt					Williams '86 MCB 6, 3807
393	apoLipoprotein B, hu	128		ccgcagctggcgATGg					Protter '86 PNAS 83, 1467
394	apoLipoprotein C-II, hu	38	*	tctctggacactATGg					Wei '85 JBC 260, 15211
395	apoLipoprotein E, rat	65	*	acaactgggaagATGa					Fung '86 JBC 261, 13777
396	Luteinizing hormone(β), rat	7		atcaagaATGg					Jameson '84 JBC 259,15474
	Lymphokines:								
397	CSF-1, hu macrophage	178		ccagctgcccgtATGa					Ladner '87 EMBO 6, 2693
398	GM-CSF, mu	35		gtcctgaggaggATGt					Miyatake '85 EMBO 4,2561
399	multi-CSF, hu (IL-3)	>38		gccgatccaaacATGa					Dorssers '87 Gene 55,115
400	" mu	29		cagaacgagacaATGg					Miyatake '85 PNAS 82, 316
401	deleted								
402	Interleukin-1α, hu	>45		aaagaagtcaagATGg					March '85 Nature 315,641
403	Interleukin-1β, hu	87	*	tctgaagcagccATGg					Clark '86 NAR 14, 7897
404	BSF-1, hu (IL-4)	>63		cgacacctattaATGg					Yokota '86 PNAS 83,5894
405	BSF-1, mu (20K)	63		acagagctattgATGg					Otsuka '87 NAR 15, 333
406	BSF-2, hu			aggagcccagctATGa					Hirano '86 Nature 324, 73

No.					w	w/t	s	s/t	
407	Lymphotoxin (TNF-β) hu	>79		ttggttctccccATGa					Gray '84 Nature 312,721
408	Lysophospholipase, rat	21		cagacactcactATGg					Han '87 Biochem 26, 1617
409	Lysozyme, ch	29		gacactggcaacATGa					Jung '80 PNAS 77, 5759
410	α_2-Macroglobulin, hu	>43		tctttctgcaacATGg					Kan '85 PNAS 82, 2282
411	α_2-Macroglobulin, rat	>63		.cctttccgcagcATGg					Gehring '87 JBC 262, 446
412	Malate dehydrogenase, mu	>50		cccgccctagccATGc					Joh '87 Biochem 26, 2515
413	Malic enzyme, mu	>65		ccggtgccagccATGg					Bagchi '87 JBC 262, 1558
414	Malic enzyme, rat			acggtgctggccATGg					Magnuson '86 JBC 261, 1183
415	Mn superoxide dismutase mu	>55		taaacctcaataATGt					Hallewell '86 NAR 14,9539
416	Melanoma Ag p97, hu	>60		cccgacggcgccATGc					Rose '86 PNAS 83, 1261
417	Menadione reductase, rat	>74		acttctggagccATGg					Robertson '86 JBC 261,15794
418	Metallothionein-I_A, hu	73		ccgcggctcgaaATGg					Richards '84 Cell 37, 263
419	" -I_B, hu	69		cttggctccacaATGg					Heguy '86 MCB 6, 2149
420	" -I_F, hu	71		cctcggcttgcaATGg					Varshney '86 MCB 6, 26
421	" -II, hu	69		cttcagctcgccATGg					Karin '82 Nature 299, 797
422	" -Ia, sh	72		cttttoctccaaATGg					Peterson '86 EJB 160, 579
423	α_1Microglobulin, hu	>72		gagcccatagccATGa					Traboni '86 NAR 14, 6340
424	β_2Microglobulin, mu	>52		tcagtcgtcagcATGg					Daniel '83 EMBO 2, 1061
425	Mullerian inhibiting subst.	10		agcacccacgATGc					Cate '86 Cell 45, 685
426	Multidrug resistance, hu	140		cgcgaggtcgggATGg					Ueda '87 JBC 262, 505
427	Mx protein, mu	>213		gagagccagacgATGg				1	Staeheli '86 Cell 44, 147
428	Myelin basic protein, mu	47		ggcttggatgtgATGg		1			Takahashi '85 Cell 42, 139
429	Myelin P2 protein, mu	>44		aaggtttacaaaATGt					Bernlohr '84 PNAS 81,5468
430	Myelin P_0(peripheral), rat	>31		cctaccccagctATGg					Lemke '85 Cell 40, 501
431	Myelin-assoc."MAG", rat	>130		ttgctggacaagATGa					Arquint '87 PNAS 84, 600
432	Myeloperoxidase, hu	>163		aggagaagagagATGg		1			Morishita '87 JBC 262,3844
433	Myoglobin, hu	70		tcagactgcgccATGg					Weller '86 MCB 6, 4539
434	Myoglobin, mu	55		ttagaagccaccATGg					Blanchetot'86 EJB 159,469
	Myosins:								
435	H-chain, rat embry. skel.	90	**	tcagccaacactATGa					Strehler '86 JMB 190, 291
436	H-chain(fast), ch. adult	60	*	gtgagcgcagccATGg					Gulick '85 JBC 260, 14513
437	H-chain(fast), ch.embry.	101	**	taaacagcgacgATGg					"
438	L-chain 1, mu	125		cttttaatcaaaATGg					Robert '84 Cell 39, 129
439	L-chain 3, mu	94		tagaactccatcATGt					"
440	L-chain 2, rat skel.	56		aggatctaagacATGg					Nudel '84 NAR 12, 7175
441	L-chain 1, ch skel.	123		aagcaacacaaaATGg					Nabeshima'84 Nat 308,333
442	L-chain 3, ch skel.	71		caactctcaatcATGt					" '82 NAR 10, 6099
443	L-chain 2A, ch cardiac			ctctgcgaagacATGg					Winter '85 JBC 260, 4478
444	Neurofilament p68, mu			ccggccgccaccATGa					Lewis '86 MCB 6, 1529
445	Neuroleukin, mu	>52		gggtccctcgccATGg					Gurney '86 Sci 234, 566
446	Neural cell adhesion, ch	>215		ccgccggctgcgATGc					Edelman '87 Sci 236, 799
447	Neural cell adhesion, mu	161		cggcagtttacaATGc					Barthels '87 EMBO 6, 907
448	Neuropeptide Y, hu	86	*	gcgccagccaccATGc					Minth '86 JBC 261, 11974
449	Nerve GF (α) mu	42		acacctgttaccATGt					Evans '85 EMBO 4, 133
450ℓ	" " (β) submax. gland	99	*	ctcctagtgaacATGc					Selby '87 MCB 7, 3057
451	" " (γ) mu	42		acacctgtcaccATGt					Evans '85 EMBO 4, 133
452	Nuclear prot. N1/N2, Xp	>64		gggttgctgatcATGg					Franke '86 EMBO 5, 3547
453	Nucleoplasmin, Xp	>113		tatctacgtgacATGg		2			Dingwall'87 EMBO 6, 69

No.				w	w/t	s	s/t	
	proto-Oncogenes:							
454	c-abl, mu, type I mRNA	>93	ggccacgggaccATGt					Ben-Neriah'86 Cell 44,577
	type IV mRNA	>200	tattattgctttATGg		1		1	
455	c-bcl-2, hu, 5.5 kb mRNA	>1000	cctctgggaaggATGg		6		2	Croce '86 PNAS 83, 5214
	3.5 kb mRNA	>150					1	"
456	c-bcr, hu	>534	gccggccgcgccATGg		1		1	Adams '87 EMBO 6, 115
457	c-erb-A, hu	>300	acccccaacagtATGa		4		2	Evans '86 Nature 324, 641
458	c-erb-A, ch	>288 *	gaattgcggtgaATGg		2		1	Sap '86 Nature 324, 635
459	c-fes/fps, fe	*	gcggacggcactATGg					Roebroek '87 JV 61, 2009
460	c-fms, hu	>300	cccaccgaggccATGg					Coussens '86 Nature 320, 277
461	c-hst, hu	>238	cctcgggccgggATGt		1			Taira '87 PNAS 84, 2980
462	c-int-1, mu	184 +	gccaggcaggccATGg				1	Varmus '85 MCB 5, 3337
463	c-int-2, mu	>326	cgcgatgccgggATGg	1				Moore '86 EMBO 5, 919
464	pp56-LSTRA, mu	>193	ccgggagggatcATGg					Sefton '86 Nature 319,682
465	c-lyn, hu	297	cgagcgggaaatATGg					Yamanashi '87 MCB 7, 237
466	c-mos, mu, ovarian mRNA	~70	tctgagggtgtaATGc		2			Propst '87 MCB 7, 1629
	" testicular mRNA	280			4			
467	c-myb, hu	>113	gcccgccgcgccATGg					Majello '86 PNAS 83, 9636
468	c-myb, mu	200 to 680	gcccgcctcgccATGg		±			Watson '87 EMBO 6, 1643
469	c-myc, hu	400, 570*	cctcccgcgacgATGc				±	Saito '83 PNAS 80, 7476
470	c-myc, fe	400, 587*	gcaggcgccgcgATGc					Stewart '86 Virol 154,121
471	c-neu, hu (HER2)	178 +	gcagtgagcaccATGg				1	Tal '87 MCB 7, 2597
472	c-pim-1, mu	~400	ctggaggtggggATGc					Selton '86 Cell 46, 603
473	c-raf-1, hu	>129 *	taagctgcatcaATGg		1			Bonner '86 NAR 14, 1009
474	A-raf-1, hu	>194	atctaaggctccATGg		1		1	Beck '87 NAR 15, 595
475	c-ral, simian		ctgtgacacgagATGg					Chardin '86 EMBO 5, 2203
476	c-Ha-ras-1, hu	69 to 332 *	ccctgaggagcgATGa	(1)	(2)			Honkawa '87 MCB 7, 2933
477	c-Ha-ras-1, rat	~175 *	cctgtagaagcgATGa	1				Damante '87 PNAS 84, 774
478	c-Ki-ras, mu	200 to 250 *	ggcctgctgaaaATGa					Hoffman '87 MCB 7,2592
479	N-ras, hu	~245	tgctggtgtgaaATGa		2			Hall '85 NAR 13, 5255
480	R-ras, hu	65	agcggtggcgacATGa					Lowe '87 Cell 48, 137
481	rho (ras-related)	>159	gttgcctgagcaATGg					Yeramian '87 NAR 15, 1869
482	c-sis, hu (PDGF2)	1022	cccggagtcggcATGa		2		1	Rao '86 PNAS 83, 2392
483	c-src, ch	>100 *	cagcccaccaccATGq					Takeya '83 Cell 32, 881
484	c-src, Xp	>58	caacaggacaagATGg		1			Steele '85 NAR 13, 1747
485	c-syn, hu ("c-slk")	>589	ggaatttagataATGg		1		2	Semba '86 PNAS 83, 5459
486	c-yes, hu	>162	gcagatttgataATGg					Sukegawa '87 MCB 7, 41
487	p53, hu	138, 230*	cgggtcactgccATGg		(2)		(1)	Lamb '86 MCB 6, 1379
488	preproOpiomelanocortin,bo	129 *	cctgcctggaagATGc					Nakanishi '81 EJB 115,429
489	preproOpiomelanocortin, Xp	62 *	tccagtcctgaaATGt					Martens '87 EJB 165, 467
490	Ornithine ATase, hu	>54	ttgaaggacacaATGt					Inana '86 PNAS 83, 1203
491	Ornithine ATase, rat	>50	aggacccacacaATGc					Mueckler '85 JBC 260,12993
492	Ornithine decarboxylase	>300	acatcgagaaccATGa		1			Gupta '85 JBC 260, 2941
493	OTCase, mu	136	agcaaaaagaagATGc					Veres '86 JBC 261, 7588
494	Ovalbumin, ch	64	tcagagttcaccATGg					McReynolds '78 Nat 273,723
495	Ovoinhibitor, ch		aggtgctctgccATGa					Scott '87 JBC 262, 5899
496	ovomucoid, ch	53	cagtacctcaccATGg					Catterall'80 JCellB 87,480
497	3-Oxoacyl-CoA thiolase, rat	>100	ctgagcttcgtcATGg				7	Arakawa '87 EMBO 6, 1361
498	preproOxytocin, bo	33	cgcgtctgcaccATGg					Ruppert '84 Nature 308,554
499	preproOxytocin, rat	40	aacaccaacgccATGg					Ivell '84 PNAS 81, 2006
500	Pancreatic polypeptide, hu	>50 *	tctggactccggATGg					Leiter '85 JBC 260, 13013
501	Parathyroid hormone, hu	>70 *	ttgtatgtgaagATGa		1			Vasicek '83 PNAS 80, 2127
502	Parotid secretory protein mu	55 *	agcaaaccaaagATGt					Poulsen '86 EMBO 5, 1891
503	Pepsinogen, hu	54	ccgggaagaaccATGa					Sogawa '83 JBC 258, 5306
504	Pepsinogen, rat	>60	caaaccggcatcATGa					Ichihara '86 EJB 161, 7
505	Peroxi. enoyl-CoA hydratase	24	taccttgagaaaATGg					Ishii '87 JBC 262, 8144

No.				w	w/t	s	s/t	
506	Phenylalanine hydroxylase	>222	cggggagccagcATGt					Kwok '85 Biochem 24, 556
507	Phosphate carrier prot. bo	>62	cttagggagaagATGt					Runswick '87 EMBO 6, 1367
508	Phosphodiesterase, cyclic	>59	ttcttccgcaaaATGt					Kurihara '87 JBC 262,3256
509	PEP carboxykinase, rat	143 *	accattgcaagaATGc					Beale '85 JBC 260, 10748
510	PEP carboxykinase, ch	166, 246 *	gcagctgcagtaATGg		(1)		(1)	Cook '86 PNAS 83;7583
511	PGK-1 (X-linked), hu	94	tgtatttccaaaATGt					Riggs '84 Gene 32, 409
512	PGK-2, mu testicular	>20	cataccatcaagATGg					Boer '87 MCB 7,3107
513	PGK, mu X-linked		ggtcttgccaaaATGt					Mori '86 Gene 45, 275
514	γ-Phosphor-kinase, mu		atccacgtgaccATGa					Caskey '87 PNAS 84, 2886
515	Phosphorylase, purine, hu	>109	gtctgcgagaccATGg					Williams '84 NAR 12, 5779
516	Pituitary hormones, α, hu	100 *	gaaaggagcgccATGg					Fiddes '81 JMAG 1,3
517	" : α, mu	100	tgcagaagagctATGg					Chin '81 PNAS 78, 5329
518	Plasma cell glycoprotein	>111	cagagcggggcgATGg					vanDriel '87 JBC 262,4882
519	u-Plasminogen activator hu	119	gacctcgccaccATGa					Riccio '85 NAR 13, 2759
520	PDGF, A-chain, hu	>387	cctcgggacgcgATGa			1	1	Betsholtz '86 Nat 320,695
521	Platelet factor 4, rat	73	caccctcttgacATGa					Doi '87 MCB 7,898
522m	Polymerase-β(DNA), rat	51	gtccccggcaccATGa					Yamaguchi '87 MCB 7.2012
523	Polymerase II (RNA), mu	406	gcctgcctcgccATGc					Ahearn '87 JBC 262,10695
524	Poly(A) binding protein, hu	>502	agccgtgccgagATGa					Grange '87 NAR 15, 4771
525	Porphobilinogen deaminase	>83	aacagcccaaagATGa					Raich '86 NAR 14, 5955
526	Prealbumin, hu	26	attcttggcaggATGg					Sasaki '85 Gene 37, 191
	Prolactin growth hormone family:							
527	prePlacental lactogen, hu	62	cacctagtggcaATGg					Saunders '83 JBC 258, 3787
528	prePlacental lactogen, mu	>59	aactcctcagagATGa					Jackson '86 PNAS 83, 8496
529	Proliferin, mu	68	gactctgcagagATGc					Linzer '87 EMBO 6, 2281
530	Prolactin, hu	57	acgatcacgaacATGa					Truong '84 EMBO 3, 429
531	Prolactin, rat	>51	gtggtcatcaccATGa					Cooke '80 JBC 255, 6502
532	Prolactin, bo	>67	atcatcaccaccATGg					Sasavage '82 JBC 257,678
533	Proline-rich (acidic).ha	33	gcctcctccaagATGc					Ann '87 JBC 262, 3958
534	" " (glycosylated)	>34	gcctccagcgagATGc					Maeda '85 JBC 260, 11123
535	Prostatic BP, C2 chain, rat	41	aaactgagcaccATGa					Delaey '87 NAR 15, 1627
536	Protamine 1, mu	92	caagccagcaccATGg					Peschon '87 PNAS 84,5316
537	Protamine, trout	14	ccatcaatcacaATGc					Gregory '82 NAR 10, 7581
	Proteases: see also 108,109,114,126,363,503-4,519							
538	-batroxobin, snake	179	agagttgaagctATGg					Itoh '87 JBC 262, 3132
539	-ser protease, mu adipose	19	cctgctgtcagaATGc					Min '86 NAR 14, 8879
540	-Ca^{2+}protease, hu	105-155 *	tgagtcgcagccATGt					Miyake '86 NAR 14, 8805
541	-Ca^{2+}protease, ra	>150	tgagccgcagccATGt	2				Emori '86 JBC 261, 9472
542	-Ca^{2+}protease, ch	37	cagtacgcagctATGa					Ohno '84 Nature 312, 566
543	-cys protease, mu	>59	ggtgtttgaaccATGa					Portnoy '86 JBC 261,14697
544	-ser protease, EGF binding		acacctgttaccATGt					Lundgren '84 JBC 259,7780
545	-mast cell protease, rat	35	accactggcacaATGc					Benfey '87 JBC 262, 5377
546	-ser p'ase, cytotoxic T cells	>111	cttccggggaagATGa					Brunet '86 Nature 322,268
547	PDI (disulphide isomerase), rat		ccgacgtccgacATGc					Edman '85 Nature 317, 267
548m	Proteoglycan 19 (chondroitin)		gagctggtcaggATGc					Bourdon '86 JBC 261,12534
549	" 38K core protein, hu	>91	atgagataaatcATGa		1			Krusius '86 PNAS 83, 7683
550	Proteolipid protein, mu	162	agtgccaaagacATGg					Hudson '87 PNAS 84, 1454
551	Pulmonary surfactant, hu protein-A	53	ggacccagagccATGt					White '85 Nature 317, 361
552	Pyruvate kinase, ch	80 *	actccagtaaccATGt					Lonberg '83 PNAS 80,3661
553	Quinone reductase, rat	>113	ttcaactatgccATGa			1	1	Bayney '87 JBC 262, 572

No.				w w/t	s s/t	
554	snRNP-B" antigen (U2), hu	>125	tttaacacaaacATGg			Habets '87 PNAS 84,2421
555	RNP-C protein, hu	>122	ccatcaaacacgATGg			Swanson '87 MCB 7, 1731
556	U1-RNA-associated 70K, hu	>680	ggcgagacgaagATGg	5	4	Theissen '86 EMBO 5, 3209
	Receptors:					
557	β_2-adrenergic, hu	190	agactgcgcgccATGg	1		Kobilka '87 PNAS 84, 46
558	β-adrenergic, turkey	69	cgccccgcagccATGg			Yarden '86 PNAS 83, 6795
559	-for asialo-GP, L1 chain	65	cccagtgctatcATGa			Leung '85 JBC 260,12523
560	-for " L2 chain, rat	>153	cctagggccatcATGg			McPhaul '87 MCB 7, 1841
561	-for EGF, hu (HER1)	100	cggggagcagcgATGc			Ishii '85 PNAS 82,4920
562	-for estrogen, hu	232	cggccacggaccATGa	1		Green '86 Nature 320, 134
563	-GPIIIa(rel. to integrin), hu		gaggcggacgagATGc			Fitzgerald '87 JBC 262,3936
564	-for IgE, hu	213	agcaggaccgccATGg			Ikuta '87 PNAS 84, 819
565	-for IgA & IgM(epithel)	>123	cagccaccagccATGg			Mostov '84 Nature 308, 37
566	-for insulin, hu		gctcccgcagccATGg			Ullrich '85 Nature 313, 756
567	-for insulin-like GF-I, hu		caaataaaaggaATGa			Ullrich '86 EMBO 5, 2503
568	-for interleukin-2(α), hu	159, 217	agggtcaggaagATGg	2		Leonard '85 Sci 230, 633
569	-for LDL, hu	~80	gaggctgcgagcATGg			Sudhof '85 Sci 228,815
570	-for nerve growth fact. hu	>113	gggaggcgggcgATGg			Johnson '86 Cell 47, 545
571	-for PDGF, mu	>138	agcccggacaccATGg			Yarden '86 Nature 323, 226
572	-for progesterone, ra	>125	gttcaggtcgacATGa			Loosfelt '86 PNAS 83,9045
573	-for SRP		cctgctgccgccATGc			Lauffer '85 Nature 318,334
574	preproRelaxin, rat	~60	gcccagaccggaATGt			Hudson '81 Nature 291,127
575	preproRenin, hu	44	actgagggaagcATGg			Fukamizu '86 Gene 49, 139
576	Retinol-Binding protein, hu	86 *	ttcctgggcaagATGa			Cortese '85 EMBO 4, 1981
577	Retinol-BP, rat	>94	tctgtccccaaaATGc			Sherman '87 PNAS 84, 3209
578	Retinol-BP-II, rat	>55	gaggccgccatcATGa			Li '86 PNAS 83, 5779
579	Retinol-BP, Xp	>40	ttgtgaaagaagATGg			McKearin '87 JBC 262,4939
580	Rhodopsin, bo	96	agggccgcagccAUGa			Nathans '83 Cell 34, 807
581	Ribonuclease, panc. rat	75	agcaaagccactATGg			MacDonald'82 JBC 257,14582
582	Ribonucl. reduct. M1, mu	>242	cttctagcggcgATGc			Caras '85 JBC 260, 7015
583	" " M2, mu	>62	ccctcgttcgccATGc			Thelander '86 MCB 6, 3433
584	Ribophorin II, hu	>268	ctgctcggaggaATGg	1		Crimaudo '87 EMBO 6, 75
	Ribosomal proteins:					
585	rp S14, hu	40 *	cgacgtgcagaaATGg			Rhoads '86 MCB 6, 2774
586	rp S16, mu	52	gtgctcggagctATGc			Wagner '85 MCB 5, 3560
587	rp S19, Xp	>46	atagccggcaagATGa			Amaldi '82 Gene 17,311
588	rp L1, Xp	40	agcagcgaggagATGg			Loreni '85 EMBO 4, 3483
589	rp L14, Xp	39	acagccgccatcATGg			Beccari '87 NAR 15,1870
590	rp L27, mu		tctgccaccgctATGc			Belhumeur '87 NAR 15, 1019
591	rp L30, mu	32 *	taaggcaggaagATGg			Wiedemann '84 MCB 4,2518
592	rp L31, rat	>25	gggcccggcagaATGg			Tanaka '87 EJB 162, 45
593	rp L32, mu	51 *	tcaaaaggcatcATGg			Dudov '84 Cell 37, 457
594	rp L44, hu	>83	cctgctgcaaagATGg			Davies '86 Gene 45, 183
595	RSV-induced 9E3 protein, ch	77	acactcctaaccATGa			Sugano '87 Cell 49,321
596	Scrapie PrP27-30, ha	90 *	agatcagccatcATGg			Basler '86 Cell 46, 417
597	Secretogranin I, hu	>112	ccgagcggggccATGc			Benedum '87 EMBO 6, 1203
598	Seminal vscl prot IV, rat	22	ttttctggcaagATGa			Kandala '83 NAR 11, 3169
599	Sodium ch.prot I, rat	>251	caggatgacaagATGg	5	2	Noda '86 Nature 320, 188
600	Somatostatin-I, hu	105	cgcggcgccgagATGc			Shen '84 Sci 224,168
601	" rat	100	gaggcaggggagATGc			Dixon '84 JBC 259,11798
602	" -II, anglfsh	>59	ccagcagacagtATGc			Hobart '80 Nature 288,137
603	" -22, catfish		gctaccaagaagATGt			Dixon '82 PNAS 79,5152
604	Sorcin/V19, ha		gtagtcttcaccATGg			Borst '86 EMBO 5, 3201
605	SPARC, mu embry. endoderm	90	gttcccagcatcATGa			Mason '86 EMBO 5, 1465
606	Stearyl CoA desaturase, rat	>102	ccgacagccacgATGc			Thiede '86 JBC 261, 13230

No.					w	w/t	s	s/t	
607	proSucrase-isomaltase, ra			caatgaaataagATGg		1			Hunziker '86 Cell 46, 227
608	2-5A Synthetase, mu			tccagacttagcATGg					Ichii '86 NAR 14, 10117
609	Synthetase, his-tRNA, ha	77		ttggcagccaggATGg					Tsui '87 NAR 15, 3349
610	t complex protein 1, mu	>60		cgtttcctgaagATGg					Willison '86 Cell 44, 727
	T-cell antigen receptor:								
611	Ti α-chain, hu	>170		cactgctcagccATGc					Yanagi '85 PNAS 82, 3430
612	Ti β-chain, hu	52		tctcactctgccATGg		?			Smith '87 NAR 15, 4991
613	Ti γ-chain (CD3), hu	>140		agagaggaaggcATGc					Littman '87 Nature 326, 85
614	T- α-chain, mu	>55		caaggctcagccATGc					Saito '84 Nature 312, 36
615	T- V_β chain (3H.25), mu			ttctaagccaccATGg					Goverman '85 Cell 40, 859
616	T- V_β chain (5.1), mu			ctgagaggaagcATGt					Chou '87 PNAS 84, 1992
617	T- V_γ chain, mu			ctacagcagaccATGa					Garman '86 Cell 45, 733
618	T3-antigen, γ-chain, hu			acagagactgacATGg					Krissansen '86 EMBO 5, 1799
619	T3-antigen, δ-chain, hu	95		ttccgctgcgagATGg	1				Tunnacliffe '86 EMBO 5,1245
620	T3-antigen, ε-chain, hu			catgaaacaaagATGc	1				Gold '86 Nature 321. 431
	T-cell differentiation antigens:								
621	CD2 (human T11)			ccaacccctaagATGa					Sayre '87 PNAS 84, 2941
622	CD4 (human T4)	>75		ggcaaggccacaATGa					Maddon '85 Cell 42, 93
623	CD4 (rat)	>53		aagcaggccaccATGt					Clark '87 PNAS 84, 1649
624	CD8 (human Leu-2/T8)	>115		ggggagcgcgtcATGg					Sukhatme '85 Cell 40, 591
625	CD8, α-chain (Lyt-2, mu)	333		ggagagcacaccATGg					Nakauchi '87 NAR 15, 4337
626	CD8, 37K-chain (rat)			aagagcgccaagATGc					Johnson '86 Nature 323, 74
	Other T-cell proteins:								
627	cytotoxic pT 49 protein, mu			cgcactgcaaggATGa					Koyama '87 PNAS 84, 1609
628	16K MAL protein, hu	>55		cagcacgccgtcATGg					Alonso '87 PNAS 84, 1997
629	preLy-6, mu	90		ccttctctgaggATGg					LeClair '86 EMBO 5, 3227
630	IgE binding factor (soluble)	>93		gtaaagtggaaaATGg					Martens '85 PNAS 82, 2460
631	Tachykinin (neuromedinK) bo	144	*	tccacaggcatcATGc	±				Kotani '86 PNAS 83, 7074
632	Tachykinin (substnce P), rat	99	*	gcaaaatccaacATGa					Krause '87 PNAS 84, 881
633	Thrombospondin, hu	>120		aacagctccaccATGg					Dixit '86 PNAS 83, 5449
634	Thy-1, mu	78	*	actcttggcaccATGa					Giguero '85 EMBO 4, 2017
635	Thy-1-related MRC OX-2, rat			cccagagcaaggATGg					Clark '85 EMBO 4, 113
636	Thymidine kinase, hu	60		cccggaggcgcaATGa					Kreidberg '86 MCB 6, 2903
637	" " ha			cgcacagccgccATGa					Lewis '86 MCB 6, 1998
638	" " ch			agcggcgcgaacATGa					Merrill '84 MCB 4, 1769
639	Thymidylate synthase, hu	>90		gcccgccgcgccATGc					Takeishi '85 NAR 13, 2035
640	" " mu	24, 34, 51		gactgctccgttATGc					Deng '86 JBC 261, 16000
641	proThymosin-α, hu	>177		gcgtgccccaccATGt					Berger '86 PNAS 83, 9403
642	Thymosin-β_4, rat			cttccagcaaccATGt					Horecker '84 PNAS 81, 2295
643	Thyroglobulin, hu	41		agggccaggaaaATGg					Christophe '85 NAR 13, 5127
644	Thyroglobulin, bo	41		aaggctcccaagATGg					Mercken '85 Nature 316, 647
645	Thyrotropin, β-subunit, mu	>89		gttgttcaaagcATGa					Gurr '83 PNAS 80, 2122
646	" -releasing hormone Xp	>109		acagcaggaaagATGg					Richter '84 EMBO 3, 617
647	Thyroxine-binding globulin, hu			cttccttccaaaATGt					Flink '86 PNAS 83, 7708
648	Transferrin, hu	50		cgcacccggaagATGa					Lucero '86 NAR 14, 8692
649	TGF-α, hu			cccgcccgtaaaATGg					Derynck '84 Cell 38, 287
650	TGF-β1, hu	>841		gccgcctcccccATGc					Derynck '85 Nature 316, 701
651	Transin, rat			aagccagtggaaATGa					Breathnach '87 NAR 15, 1139
652	Transin-2, rat			aaggctgtctctATGg					"
	Translation factors:								
653	Elongation factor 1α, hu	>53		ctaaaagccaaaATGg					Brands '86 EJB 155, 167
654	Elongation factor 2, ha	>77		ccatccgccactATGg					Kohno '86 PNAS 83, 4978
655	eIF2, rat			atacacttcagaATGc					Ernst '87 JBC 262, 1206
656	cap binding protein, hu	18		gatcgatctaagATGg					Rychlik '87 PNAS 84, 945

No.					w w/t	s s/t	
657	Transcription fact. TFIIIA	50		gctgaaggagagATGg			Tso '86 NAR 14, 2187
658	Triose-P-isomerase, hu	34		ctcggctcggccATGg			Brown '85 MCB 5, 1694
659	" ch	52		gtcgcctccgccATGg			Straus '85 MCB 5, 3497
660	Tropomyosin, hu fibroblasts	118		ccaccgcaggccATGg			MacLeod '85 PNAS 82, 7835
661	TM30$_{pl}$ " " "	>50		gcgctccgcgccATGg			MacLeod '87 JMB 194, 1
662	TM-1, rat fibroblasts	>61		cccaccgcagccATGg			Helfman '85 JBC 260,14440
663	α-Tropomyosin, rat muscle	78		gccaccgccaccATGg			Ruiz-Opazo'87 JBC 262,4755
664	Troponin-I, fast muscle	82	*	atctaaagcaagATGt			Baldwin '85 PNAS 82,8080
665	Troponin-T, rat	79	*	cccaccgccactATGt			Breitbart '86 JMB 188,313
666	Troponin-C, slow muscle, ch			ccctgcccggccATGg			Putkey '87 MCB 7, 1549
667	Trypsinogen, anionic, ca	14		acttctgccatcATGa			Pinsky '85 MCB 5, 2669
668	" cationic, ca	29		cagggagcaaccATGa			"
669	α-Tubulin I, CHO	>100		cccgtagctaccATGc			Elliott '86 MCB 6, 906
670	α-Tubulin II, CHO	>100		aaagcagcaaccATGc			"
671	α-Tubulin III, CHO	>130		ttcctagacaccATGc			"
672[n]	α-Tubulin, hu	211		tcccgggaaaacATGc	1		Hall '85 NAR 13, 207
673	β-Tubulin, hu	72		gccgccgccatcATGa			Lewis '85 JMB 182, 11
674	β-Tubulin, hu	159		taaattttaaccATGa			Lee '83 Cell 33, 477
675	β-Tubulin, ch	>87		gacaccggcatcATGc			Cleveland'81 Nat 289,650
676[o]	β3-Tubulin, ch	40 to 50		gccgaagccatcATGa		±	Sullivan '86 JBC 261,13317
677	β4-Tubulin, ch	74		tccggccgcaccATGa			Sullivan '86 MCB 6, 4409
678	β5-Tubulin, ch	48		cgggacagcgccATGa			"
679	Tyrosinase, mu	>174		gcttcgagaagaATGa	2		Shibahara '86 NAR 14,2413
680	Tyr aminotransferase, rat	97	*	gcttcgagaggcATGg			Grange '85 JMB 184, 347
681	Tyr hydroxylase, hu	~30		ccacactgagccATGc			Grima '87 Nature 326, 707
682	Tyr hydroxylase, rat	35		ccagcttgcactATGc			Harrington'87 NAR 15,2363
683	Ubiquitin, hu	100	*	taacaggtcaaaATGc			Baker '87 NAR 15, 443
684	Ubiquitin, ch	63	*	ggagacgtaaacATGc			Bond '86 MCB 6, 4602
685	UDP-glucuronosyl-tr'ase, hu			cattgcatcaggATGt			Jackson '87 BiochJ. 242,581
686	" " steroid-induced, rat	>75		ttgattttaagATGc			Harding '87 NAR 15, 3936
687	" " 3-MC induced, rat	>124		ctctctgaaaggATGg	1		Iyanagi'86 JBC 261,15607
688	Uncoupling prot.(brown fat)	177		ctccgagccaagATGg			Ricquier'86 JBC 261,1487
689	major Urinary pr(MUP), mu	60		ctccctaccaaaATGa			Shahan '87 MCB 7, 1938
690	Uroporphyrinogen decarb. hu			agacagctgaccATGg			Goossens'86 JBC 261,9825
691	Urotensin-I, carp	>75		cctgtgtccagcATGa			Ishida '86 PNAS 83, 308
692	Uteroglobin, ra	47		cattctgccaccATGa			Suske '83 NAR 11, 2257
693	Valosin "precursor", po	>143		gaagcgcgcgccATGg			Koller '87 Nature 325, 542
694	Vasoactive intest. peptide	174	*	agaggcacagaaATGg			Linder '87 PNAS 84, 605
695	Vasopressin-neurophysinII	48		acccgtgccaggATGc			Ruppert '84 Nature 308, 554
696	Vimentin, ha	135		gctctccaaaccATGt			Quax '83 Cell 35,215
697	Vinculin, ch	>246		cccgctgccgccATGc			Price'87 Bioch.J.245,595
698	Vitellogenin-II, ch	13		ttcaccttcgctATGa			Geiser '83 JBC 258,9024
699	Vitellogenin, Xp	13		ttcgccatcaccATGa			Walker '83 EMBO 2, 2271
700	preWhey acidic protein, rat	33		gccgccgacaccATGc			Campbell '84 NAR 12, 8685
701[p]	preXenopsin, Xp	>62		catttggaaaggATGt			Sures '84 PNAS 81, 380

[a] A two-latter abbreviation indicates the source of each mRNA: bovine; canine; chicken; feline; gp, guinea pig; hamster; human; murine; porcine; rabbit; sheep; Xp, Xenopus.

[b] The number of nucleotides comprising the 5'-untranslated sequence is indicated. In most cases this was determined by primer extension and/or S1 mapping. An entry is marked > if the cDNA included a considerable portion of the 5'-noncoding

sequence but was probably not complete. There is no entry in this column if the cDNA included little of the leader sequence or if the 5'-end of the mRNA was not mapped on the genomic sequence. If a second form of mRNA was detected but its leader sequence not precisely mapped, the entry is marked +. An asterisk indicates that the 5'-noncoding sequence is interrupted by an intron.

[c] Upstream ATG codons are designated strong (s) or weak (w) according to context (see text). They are listed according to whether or not the reading frame established by the upstream ATG codon terminates (t) before the start of the major open reading frame. If a gene produces two major transcripts, only one of which has upstream ATG codons, the upstream ATG codons are listed in parentheses. ± means that upstream ATG codons occur in only a minor species of mRNA. If an upstream ATG codon lies very near the error-prone 5'-end of a cDNA clone and has not yet been confirmed by sequencing either the gene or a second cDNA, I have temporarily entered a question mark in this column.

[d] Bibliographic data are given in condensed form: first author, year, journal, volume, and first page.

[e] Alternative splicing produces transcripts with three different leader sequences, only one of which is listed. The second form of mRNA also has three upstream ATG codons, while the third has a single weak ATG codon upstream.

[f] Eight of the upstream ATG codons lie in the same reading frame, constituting an upstream cistron with the potential to encode a 40 amino acid peptide. In view of the pattern of codon usage, Shull and Lingrel postulate that this peptide is made.

[g] It is likely that ribosomes initiate at the first and second ATG codons in these mRNAs, producing long and short forms of the encoded polypeptide. In each case the 5'-proximal ATG codon occurs unusually close to the cap and in a suboptimal context for initiation; it is not known which of those features accounts for the "leakiness." The distance from the cap to the first ATG codon is 5 nucleotides for #143, 7 n for #297 and 3 n for #330.

[h] Unlike the human HMG-CoA reductase mRNA sequence which is entered in the table, a subset of transcripts from the corresponding hamster gene have upstream ATG codons.

[i] The sequence of the chicken insulin gene has a weak ATG codon upstream from the translational start site, but it is not known whether the 5'-end of the transcript includes that ATG codon.

[j] There is a developmentally regulated switch in the promoter for IGF-II in rats. The longer transcript introduces an upstream in-frame ATG codon, and initiation at that site would add eleven amino acids to the N-terminus of IGF-II. Only the shorter transcript is represented in the table because the functional initiator codon in the longer transcript has not been verified experimentally.

[k] α-Interferons A and D are the most highly expressed in humans and both of those mRNAs have A in position -3, as shown. Other human α-IFN genes that are expressed at lower levels have C in position -3, but it is not known if that substitution accounts for their lower expression.

[l] In the major form of NGF mRNA in mouse submaxillary glands, the indicated initiator codon is the first ATG triplet in the message. The exon that carries that ATG codon is spliced out of the major NGF transcript in other tissues and an ATG codon that lies farther downstream is thereby activated.

[m] The first cDNA that was cloned fell short of the real initiator codon, resulting in misidentification of an internal ATG codon as the translational start site.

[n] Whereas the context around the initiator codon is standard in the α- and β-tubulin sequences shown in the table, one α- and one β-tubulin mRNA have been described in which the ATG codon lies in a poor context for initiation (Cowan '83 MCB 3,1738; Lee '84 NAR 12, 5823). It is not known whether or how well those particular mRNA species are translated.

[o] The chicken β3-tubulin gene produces mRNAs with very heterogeneous 5'-ends, some of which would lack the upstream ATG codon.

[p] Entries 363 and 401 have been deleted, leaving 699 sequences on which the calculations in Tables 1 and 2 are based. This includes all published sequences to which I had access as of May 31, 1987, in which the functional initiator codon has been clearly identified. Another 110 sequences were excluded because of uncertainty about which ATG codon initiates translation; that list was made available to the editor during the review of this manuscript.

GENES & GENOMES

A CHANGING PERSPECTIVE

MAXINE SINGER

President, Carnegie Institution of Washington
Scientist Emeritus, National Institutes of Health

PAUL BERG

Willson Professor of Biochemistry
Director, Beckman Center for Molecular and Genetic Medicine,
Stanford University School of Medicine

UNIVERSITY SCIENCE BOOKS
MILL VALLEY, CALIFORNIA

University Science Books
20 Edgehill Road
Mill Valley, CA 94941
Fax: (415) 383-3167

Production manager: Mary Miller
Manuscript editor: Carol Dempster
Copy editor: Sylvia Stein Wright
Text designer: Gary Head
Jacket and cover designer: Robert Ishi
Art director: Charlene Kornberg
Indexer: Maria Coughlin
Compositor: Syntax International
Printer: R. R. Donnelley & Sons

Library of Congress Catalog Number: 90-070-698

ISBN 0-935702-17-2

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

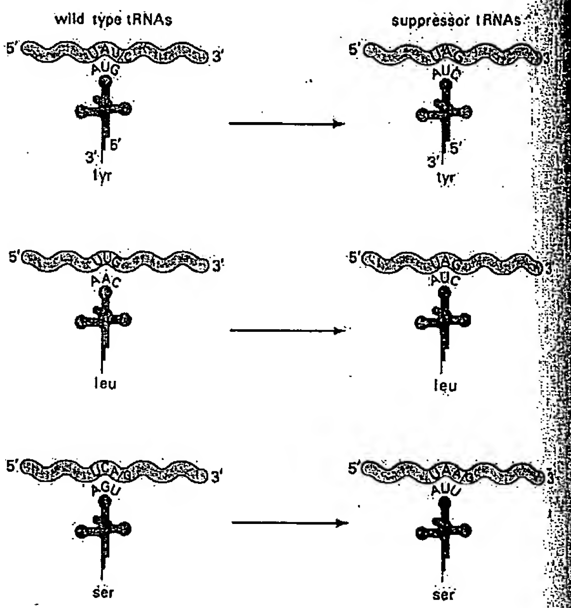


Figure 3.51
Translational suppression of terminator codons. Mutations causing a change in the anticodons of $tRNA^{Tyr}$, $tRNA^{Leu}$, or $tRNA^{Ser}$ permit translation of termination codons as amino acids.

ribosome that are involved in the codon-anticodon interactions. Also certain chemicals such as aminoglycosides (e.g., streptomycin) bind to ribosomal proteins in the 30S subunit and can alter the fidelity of translation. In these cases, there is a more widespread breakdown in the accuracy of the translation process.

3.8 mRNA Translation in Eukaryotes

Translation of eukaryotic mRNA is basically similar to that of prokaryotic mRNA. With the exceptions already noted, the genetic code is identical, and the codons are translated successively by aminoacyl-tRNAs in conjunction with ribosomes. There are, however, three notable differences imposed by certain characteristics of eukaryotic cells. First, the transcription and translation machinery in eukaryotes are physically separated, transcription occurring in the nucleus and translation in the cytoplasm. Second, the 5' and 3' ends of eukaryotic mRNAs have special structures. Third, with the exception of the mRNAs transcribed from the DNA genomes of viruses, eukaryotic mRNAs usually contain only a single protein coding sequence.

At present, we know considerably less about the structures and properties of the participants in eukaryotic translation than of their counterparts in prokaryotes. Although the same three stages—initiation, elongation, and termination—are discernable in eukaryotes, each is more complex in the number of extraribosomal protein factors that are required. In spite of the differences, protein coding sequences from prokaryotes are readily translated by the eukaryotic translation system, provided that their mRNAs possess the appropriate modifications at the 5′ and 3′ termini (Section 3.8a). Conversely, eukaryotic protein coding sequences are translated efficiently in prokaryotes, provided they contain a Shine-Delgarno sequence 5′ to the initiator AUG. This means that the translation machinery of both types of organisms can contend with the nucleotide sequence arrangements in mRNAs from whatever source.

a. Special Modifications in Eukaryotic mRNAs

Eukaryotic mRNAs transcribed from nuclear or viral DNA genes by RNA polymerase II always have a modified 5′ terminus, referred to as a "cap" (Figure 3.53). But RNAs transcribed by eukaryotic RNA polymerases I (i.e., rRNAs) and III (5S and tRNAs) are not capped and retain their original 5′-triphosphate termini. Most of the mRNA produced by animal RNA viruses is also capped, even though it is produced by virus encoded RNA transcriptases. Most uncapped mRNAs are poorly translated by eukaryotic protein synthesizing systems because of inefficient ribosome binding to the mRNA. Capping occurs at the 5′ nucleoside triphosphate and shortly after RNA transcripts are initiated, well before the transcript is completed. The details of the capping process are presented in Section 8.3c.

Eukaryotic mRNAs also contain a polyadenylate sequence at their 3′ ends. This 3′ "tail," which is 50 to 200 adenylate residues long, is not encoded in the sequence of protein coding genes, but is added posttranscriptionally after cleavage of the transcript at a specific sequence beyond the translation termination signal (Section 8.3c).

b. Initiation of Translation by Small Ribosomal Subunits at the 5′ Capped Ends of mRNAs

Just as the dissociation of 70S ribosomes is an obligatory step for initiating the translation of prokaryotic mRNA, the 80S ribosomes must be dissociated before the translation of eukaryotic mRNAs can begin. The small subunit (40S), in association with an array of accessory proteins (eIFs), one or more of which are needed to dissociate the ribosome into its component subunits, binds the special initiator met-$\text{tRNA}_\text{I}^\text{Met}$. Here, too, GTP and a special protein—eIF-2—are needed to bind the initiator aminoacyl-tRNA. In eukaryotes, however, met-$\text{tRNA}_\text{I}^\text{Met}$ is not N-formylated; but as in prokaryotes, the structure of $\text{tRNA}_\text{I}^\text{Met}$ is different from $\text{tRNA}_\text{M}^\text{Met}$. The 40S complex containing met-$\text{tRNA}_\text{I}^\text{Met}$, GTP, and a panoply of other eIFs binds to the mRNA at or near the capped 5′ end; at least one of the factors recognizes and binds to the cap structure. The small subunit moves from the capped terminus to the first AUG downstream from the 5′ end by a mechanism that is still unknown. At present, there is no requirement for a nucleotide sequence comparable to the Shine-Delgarno sequence near the AUG. However, the efficiency with which an AUG serves as an initiator codon is influenced by certain nucleotides on both sides of the AUG. This preinitiation complex com-

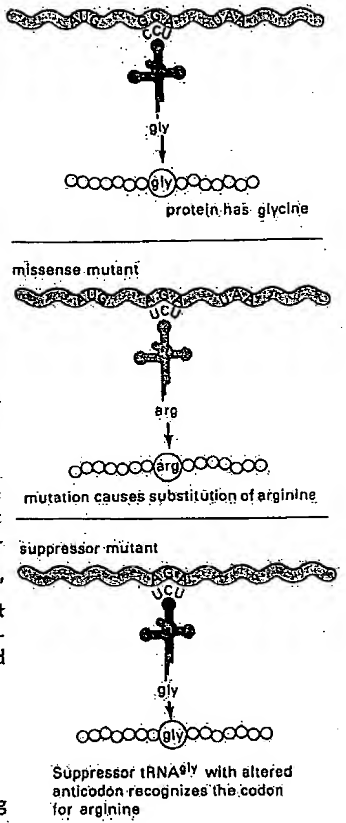

CCU
gly
gly
protein has glycine
missense mutant
UCU
arg
arg
mutation causes substitution of arginine
suppressor mutant
UCU
gly
gly
Suppressor tRNAgly with altered anticodon recognizes the codon for arginine

Figure 3.52. Translational suppression of missense mutations. In this example, a tRNA^Gly acquires an altered anticodon that permits insertion of glycine at an arginine codon, AGA.

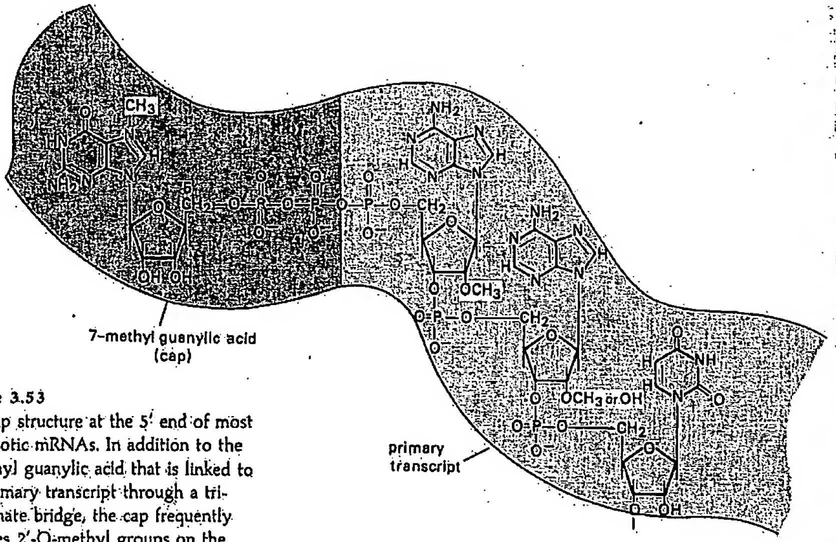


Figure 3.53
The cap structure at the 5' end of most eukaryotic mRNAs. In addition to the 7-methyl guanylic acid that is linked to the primary transcript through a triphosphate bridge, the cap frequently includes 2'-O-methyl groups on the first or first two ribose residues in the transcript.

bines with the 60S subunit in an energy- and factor-dependent reaction to create the functional initiation complex. The initiation of protein synthesis can be regulated by phosphorylation and dephosphorylation of an eIF-2.

c. Polypeptide Chain Elongation and Termination

The stepwise translation of successive codons with aminoacyl-tRNAs is not substantially different in eukaryotes and prokaryotes. GTP and the elongation factor, eEF-1, which corresponds to the prokaryotic complex of EF-Tu and EF-Ts, cycle to bring aminoacyl-tRNAs to the ribosomes. GTP and eEF-2, which is functionally analogous to prokaryotic EF-G, promote the translocation operation. Termination of translation in eukaryotes also occurs at any one of the three stop codons, causing release of free polypeptide chains, tRNA, and, very likely, the 80S ribosome from the mRNA. One factor, eRF, and GTP appear to mediate the entire series of termination events.

The biogenesis of eukaryotic mRNA is discussed in detail in Section 8.3, but here we need to stress that translation does not occur until the mature mRNA reaches the cytoplasm, whereupon initiation can occur as mentioned previously. Polysomes are formed by successive initiations, and simultaneous, multiple translations of the protein coding sequence occur as in prokaryotes. A notable difference, however, is that cellular mRNAs generally contain only a single protein coding sequence.. When mRNAs contain consecutive coding regions, as occurs with many animal DNA viruses, the downstream coding sequences are inefficiently or not translated.

DEX-0172 GE

VOLUME I GENERAL PRINCIPLES

MOLECULAR BIOLOGY OF THE GENE

FOURTH EDITION

James D. Watson COLD SPRING HARBOR LABORATORY
Nancy H. Hopkins MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Jeffrey W. Roberts CORNELL UNIVERSITY
Joan Argetsinger Steitz YALE UNIVERSITY
Alan M. Weiner YALE UNIVERSITY

The Benjamin/Cummings Publishing Company, Inc.
Menlo Park, California • Reading, Massachusetts • Don Mills, Ontario
Wokingham, U.K. • Amsterdam • Sydney • Singapore
Tokyo • Madrid • Bogota • Santiago • San Juan

Cover art is a computer-generated image of DNA interacting with the Cro repressor protein of bacteriophage λ. The image was prepared by the Graphic Systems Research Group at the IBM U.K. Scientific Centre, using coordinates courtesy of Brian W. Matthews, University of Oregon.

Editor: Jane Reece Gillen
Production Supervisor: Karen K. Gulliver
Editorial Production Supervisor: Betsy Dilernia
Cover and Interior Designer: Gary A. Head
Contributing Designers: Detta Penna, Michael Rogondino
Copy Editor: Janet Greenblatt
Art Coordinator: Pat Waldo
Art Director and Principal Artist: Georg Klatt
Contributing Artists: Joan Carol, Cyndie Clark-Huegel, Barbara Cousins, Cecile Duray-Bito, Jack Tandy, Carol Verbeek, John and Judy Waller

Library of Congress Cataloging-in-Publication Data
Molecular biology of the gene.
Rev. ed. of: Molecular biology of the gene / James D. Watson. 3rd ed. c1976.
Bibliography:
Includes index.
Contents: v. 1. General principles.
1. Molecular biology. 2. Molecular genetics.
I. Watson, James D., 1928– . [DNLM: 1. Cytogenetics.
2. Molecular Biology. QH 506 M7191]
QH506.M6627 1987 574.87'328 86-24500
ISBN 0-8053-9612-8

CDEFGHIJ-MU-898

The Benjamin/Cummings Publishing Company, Inc.
2727 Sand Hill Road
Menlo Park, California 94025

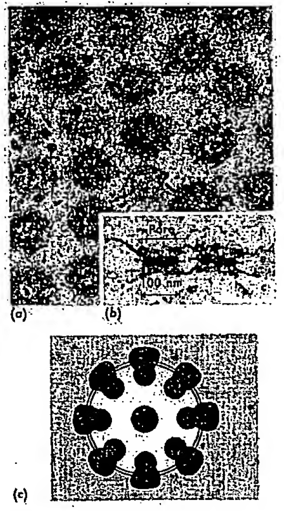


Figure 18-17
Nuclear pores. (a) Electron micrograph of part of a nuclear envelope with a number of pores. (b) Cross section through two pores (n = nucleus, e = envelope, c = cytoplasm). (c) Diagram of a pore, showing the eight pairs of granules that line it and a central granule. Although yeast does have nuclear pores, detailed electron microscopic studies have not yet been carried out. (Courtesy of K. Roberts, John Innis Institute, Norwich, England.)

rich; proteins account for about one-half their mass (as compared to one-third in bacterial ribosomes) and represent more individual polypeptide species. Why these differences exist is not known. The thing to remember is that on a structural level, there is a larger gap in ribosome size between procaryotic and eucaryotic microorganisms (e.g., between bacteria and the yeasts, *Neurospora*, or *Aspergillus*) than between the most evolutionarily divergent eucaryotic cells.

Each mature yeast rRNA chain, with the exception of that of 5S rRNA, is initially synthesized as part of a much longer rRNA precursor (37S) that is processed to give rise to the two large (25S and 17S) rRNA species and one of the small (5.8S) rRNA species. In yeast, but not in higher eucaryotes, the 5S gene is located close to the rDNA units and is part of the same repeating unit, even though it is transcribed by a different RNA polymerase and in the opposite direction. Because the rDNA units are tandemly located next to each other, intramolecular crossing over between them occasionally generates yeast cells containing small circles of rDNA, which are rapidly lost, since they do not possess centromeres. Tandem repetition, however, is important for maintaining the sequence homogeneity of the individual repeats within the larger unit by mechanisms that will be discussed in Chapter 20.

Within all eucaryotic cells, the actively transcribed rDNA units are somehow compacted into dense-appearing nucleolar bodies. In yeasts, these nucleoli are crescent shaped and closely associated with the nuclear membrane.

Caps at the 5′ Ends of Eucaryotic mRNAs[50–53]

Eucaryotic mRNAs, including those of yeast, have several important features that differentiate them from procaryotic mRNAs. First, they are modified at their 5′ ends by the addition of a guanine nucleotide. This alone would not be considered unusual if the linkage were a conventional 3′–5′ phosphodiester bond. Instead, GTP reacts with the triphosphate at the 5′ end of an mRNA chain in a 5′–5′ condensation to form the structure 3′-G-5′ppp5′-N-3′p . . . , generally known as a cap. Thus, the 5′ and 3′ ends of most eucaryotic mRNAs are terminated by ribose moieties with free 2′- and 3′-OH groups. Subsequent to cap formation, a methyl group is added to the backward guanine residue (at the 7 position of the purine ring) and often also to the 2′-OH groups of the first and/or second adjacent nucleotides (Figure 18-18).

Why do the 5′ ends of eucaryotic mRNAs need to be so blocked? One possible reason is that these caps (and apparently specific proteins that bind to them) help ribosomes attach to mRNA chains so that they start translation at the correct AUG codon. Favoring this hypothesis is the absence of sequences complementary to 18S rRNA preceding coding regions in eucaryotic mRNA molecules. Specific ribosome binding sequences analogous to those of procaryotic mRNAs thus do not exist on eucaryotic mRNA molecules. Instead, eucaryotic ribosomes search out AUG initiator codons by binding to the caps and then migrating to the closest (5′-most proximal) AUG codon to start translation. Apparent exceptions to this rule are several viral mRNAs (e.g., polio RNA) that function perfectly normally in eucaryotic cells but lack the 5′ cap structure. Their ends are blocked instead by specific proteins that perhaps substitute as positioning agents.

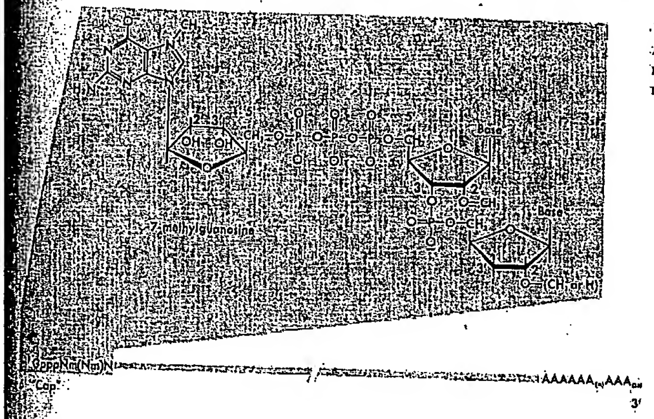


Figure 18-18
A generalized structure for eucaryotic mRNA, showing the posttranscriptional modifications at the 5′ and 3′ ends.

Yeast mRNA Molecules Code for Single (Never Multiple) Polypeptide Chains[54, 55]

A given yeast mRNA molecule, like most other eucaryotic mRNAs, carries the genetic message for only a single polypeptide chain. Unlike many bacterial mRNAs, where translation can begin and end at several sites to give rise to several independent polypeptide chains, eucaryotic mRNAs are constructed so that translation usually commences at the first AUG codon following the capped 5′ end. (Exceptions will be discussed in Chapter 24.) Thus, eucaryotic mRNAs are generally monocistronic, as opposed to the frequently polycistronic mRNAs of bacteria. Like bacterial mRNAs, however, eucaryotic mRNA molecules usually have extensive untranslated sequences before and after their protein-coding regions (leader and trailer sequences).

The inability of eucaryotic mRNAs to encode multiple proteins and thereby ensure their coordinate regulation probably explains why many eucaryotic polypeptides consist of several independent domains having related enzymatic functions. The yeast *HIS4* mRNA, for example, codes for a single protein that carries out three different enzymatic steps in histidine biosynthesis. Yet, it is rare that a single polypeptide carries out all the steps involved in a given metabolic pathway. Thus, regulatory molecules that coordinate the transcription of several functionally related mRNAs must act at multiple chromosomal sites. It is important to note that this latter feature is not unique to eucaryotes. For example, the enzymes that carry out the biosynthesis of arginine in *E. coli* are encoded by several different mRNA molecules whose expression is coordinately regulated, rather than being regulated by a single polycistronic mRNA.

Poly A at the 3′ Ends of Eucaryotic mRNAs[56–60]

Still very mysterious is the observation that most yeast mRNA molecules, like those of all other eucaryotic cells, contain relatively long stretches (about 200 residues) of poly A at their 3′ ends. These poly A

X. Related Proceedings Appendix

There are no related proceedings.